DOCUMENT RESUME

ED 165 278 EA 011 194

AUTHOR Clark, Harry L.; Thurston, Dona R.
TITLE Planning Your Staffing Needs: A Handbook for Personnel Workers.
INSTITUTION Civil Service Commission, Washington, D.C. Bureau of Policies and Standards.
PUB DATE 77
NOTE 328p.; Not available in paper copy due to marginal legibility of original document
AVAILABLE FROM Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (Stock No. 006-000-01020-2; $5.25)

EDRS PRICE MF-$0.33 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Computer Programs; *Employment Projections; Employment Trends; *Labor Turnover; Needs Assessment; *Personnel Needs; Planning
IDENTIFIERS GS810 Program; HIREST Program; Log Probability Curve; LOGPRO Program; LPFILE Program; LPTEST Program; Retention Curves; Staffing Needs Planning

ABSTRACT

Staffing needs planning, as distinct from workforce planning, is the systematic estimation of the actions personnel managers must take to provide the required workforce in an organization. This handbook focuses on the pivotal aspects of staffing needs planning, including estimation of future vacancies and analysis and projection of employee turnover. The first section of the book briefly introduces staffing needs planning and the policy issues related to it. The second section explains details of manual methods of projecting turnover losses, including the nature of the retention curve and fitting and using the log-probability curve. Section 3 presents computer methods for analyzing turnover, advancement, and hiring needs, including LOGPRO, LPFILE, LPTEST, GS810, and HIREST. The handbook is designed for use by those with limited statistical training as well as by those who are technically proficient. The computer programs are intended to be general in application, complete, compact, simple, and fully documented. Appendices provide detailed documentation on the manual methods and the computer programs, including full computer program listings, operation manuals, and technical analysis. (Author/JM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

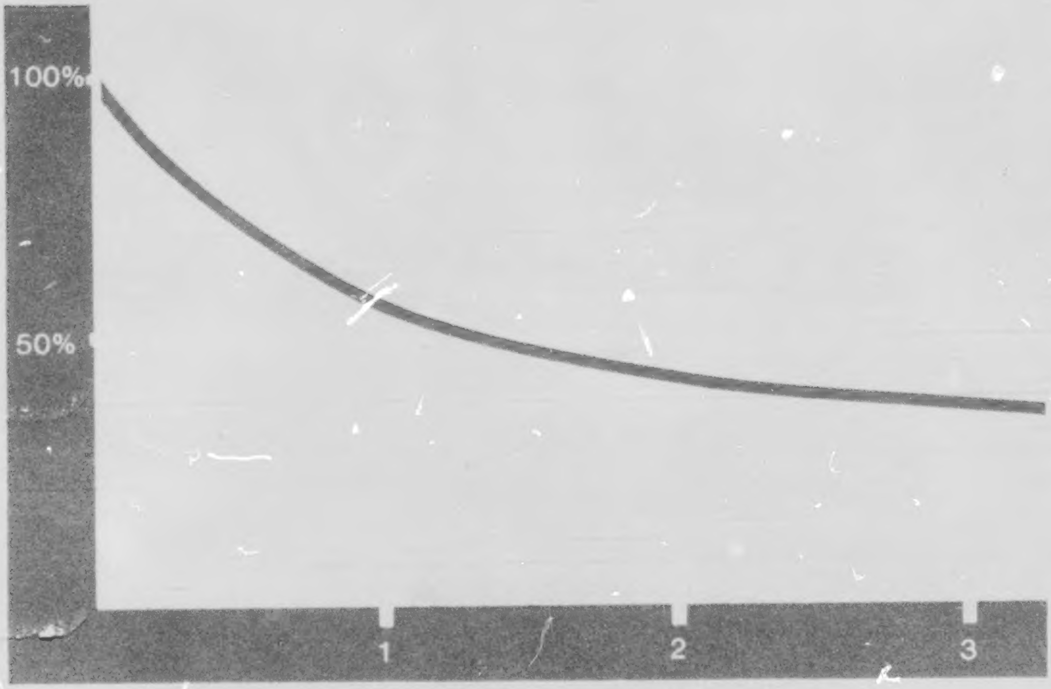
ED165278

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Planning Your Staffing Needs

A Handbook for Personnel Workers



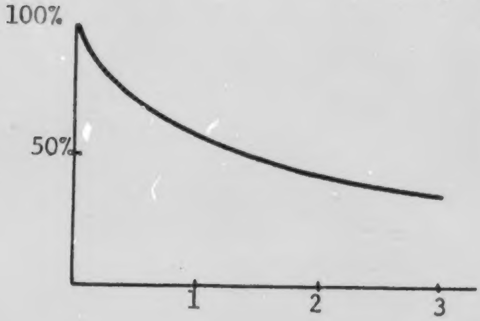
EA 011 194

U.S. Civil Service Commission
Bureau of Policies and Standards
Washington, D.C.
1977

Acknowledgement

This handbook was written by Harry L. Clark and Dona R. Thurston, Manpower Analysis Officer and Mathematician, respectively, in the Bureau of Policies and Standards' Policy Analysis and Development Division.

PLANNING YOUR STAFFING NEEDS: A HANDBOOK FOR PERSONNEL WORKERS



Bureau of Policies and Standards
United States Civil Service Commission
Washington, D. C.

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 (Paper Cover)

Stock No. 006-000-01020-2

PREFACE

This handbook is a product of the U.S. Civil Service Commission's continuing program of manpower planning research. Earlier stages of this research have been published in the former Federal Workforce Outlook and Current Federal Workforce Data publications series.

This handbook is divided into two main parts, a narrative text and a series of appendices. The text deals with three principal topics:

- Staffing needs planning policy matters (Chapter 1);
- Manual methods of projecting turnover losses (Chapters 2-5); and
- Computer methods for analyzing turnover, advancement and hiring needs (Chapters 6-10).

The appendices provide detailed documentation on the manual methods and the computer programs, including full computer program listings, operation manuals and technical analyses.

Within the text, all statistics are presented in an elementary, step-by-step manner which thoroughly explains and displays all the techniques which are utilized. This method of presentation was chosen to accommodate the wide diversity of skill levels of the audiences for which the handbook has been written. Thus the handbook can be understood and used by people with any level of quantitative knowledge: from those who have very few quantitative skills to those who are technically proficient. In addition, the handbook can be used by organizations with any type of data system: from manual recordkeeping to sophisticated computer systems.

These policies, procedures, and computer programs are offered for optional and developmental use only. They may be used in whole or in part by any agency or organization which desires to do so. In no sense, then, should their use be considered mandatory on any Federal or non-Federal office or organization.

Being developmental in nature, we would expect that these policies and programs will not necessarily be uniformly effective under all types of field conditions. We would welcome information from users as to their experience with this technology and would welcome information on the changes, or suggested changes, which users may find necessary or desirable under field conditions.

Also, the Commission stands ready to provide whatever assistance it can to organizations interested in studying or applying the techniques contained herein.

If you have any questions or information, or would like to discuss possible Commission assistance in your organization, please address:

Bureau of Policies and Standards
U.S. Civil Service Commission
Washington, D.C. 20415

TABLE OF CONTENTS

Chapter Page

PREFACE... iii

SECTION I: INTRODUCTION TO STAFFING NEEDS PLANNING

1. STAFFING NEEDS PLANNING

Introduction.. 3
Definitions and Responsibilities.................... 4
Suggested Policy Requirements....................... 6
Plan for This Handbook.............................. 10

SECTION II: MANUAL METHODS FOR ANALYZING TURNOVER

2. ANALYZING WORKGROUP LOSSES

The Technical Problem............................... 13
The Results of Turnover Research.................... 14
The Log-Probability Nomograph....................... 17
Application to Hiring Levels........................ 21
Other Applications.................................. 25
Conclusions... 26

3. ON CURVE FITTING

The General Formula for a Straight Line............. 29
The Linear Least Squares Line....................... 34
Properties of the Linear Least Squares Line......... 39

4. THE NATURE OF THE RETENTION CURVE

Data Transformations................................ 45
Some Comments About Retention Analysis.............. 54
    Data Collection................................. 54
    Why Retention is a Log-Probability Function..... 56
    Condition for Projection Validity............... 57
    Continuity of the Retention Curve............... 57

5. FITTING AND USING THE LOG-PROBABILITY CURVE

Manual Calculation of the Log-Probability Equation.... 61
Iteration and Projection of the Equation............ 64
Application to Hiring Levels........................ 67
Other Applications.................................. 69
Confidence Ranges for Projected Values.............. 73

TABLE OF CONTENTS (CONT'D)

Chapter Page

SECTION III: COMPUTER METHODS FOR ANALYZING TURNOVER, ADVANCEMENT AND HIRING NEEDS

6. STAFFING NEEDS PLANNING COMPUTER PROGRAM: LOGPRO

Analysis Functions.................................... 79
Log-Probability Analysis.............................. 81
Projection Analysis................................... 82
Summary... 83

7. STAFFING NEEDS PLANNING COMPUTER PROGRAM: LPFILE

When To Use LPFILE.................................... 87
The LPFILE Method..................................... 89
Sample Run of LPFILE.................................. 103

8. STAFFING NEEDS PLANNING COMPUTER PROGRAM: LPTEST

When To Use LPTEST.................................... 109
Statistical Differences............................... 110
The LPTEST Method..................................... 115
Sample Run of LPTEST.................................. 117

9. STAFFING NEEDS PLANNING COMPUTER PROGRAM: GS810

The Need for Analysis of Advancement.................. 125
Barriers to Analysis.................................. 126
GS810 Features.. 127
Preparation of Data................................... 130
Using GS810... 131
Conclusion.. 135

10. STAFFING NEEDS PLANNING COMPUTER PROGRAM: HIREST

First-Year Projections................................ 137
Further Years' Projections............................ 141
Projecting "Hires Needed"............................. 143
HIREST Requirements................................... 145
Conclusions... 149

11. THE ROLE OF THE ANALYST.............................. 151

APPENDICES

Page

APPENDIX A - Manual Calculation of the Log-Probability Curve 155

APPENDIX B - Some Applications of the Log-Probability Equation 169

APPENDIX C - Staffing Needs Planning Computer Program: LOGPRO

C-1 : Technical Analysis 189
C-2 : Operation Manual 201
C-3 : Program Listing 211

APPENDIX D - Staffing Needs Planning Computer Program: LPFILE

D-1 : Technical Analysis 221
D-2 : Operation Manual 227
D-3 : Program Listing 235

APPENDIX E - Staffing Needs Planning Computer Program: LPTEST

E-1 : Technical Analysis 247
E-2 : Operation Manual 253
E-3 : Program Listing 265

APPENDIX F - Staffing Needs Planning Computer Program: GS810

F-1 : Technical Analysis 279
F-2 : Operation Manual 287
F-3 : Program Listing 297
F-4 : Sample Outputs 313

APPENDIX G - Staffing Needs Planning Computer Program: HIREST

G-1 : Technical Analysis 329
G-2 : Operation Manual 337
G-3 : Program Listing 343

APPENDIX H - Running The Computer Programs 357

PLANNING YOUR STAFFING NEEDS:

A HANDBOOK FOR PERSONNEL WORKERS

CHAPTER 1

STAFFING NEEDS PLANNING

Introduction

One of the central tasks of personnel management is to help agency management acquire the numbers and kinds of workers required to carry out the organization's mission. Obviously, to do this it is first necessary to identify what the organization's staffing needs are. Only then can personnel management actions and programs to meet these needs be planned and carried out.

In the past, organizational staffing needs have been customarily determined by such means as receipts of SF-52's, Request for Personnel Action. These informed personnel officials of the existence of vacancies that needed to be filled. In some cases, personnel officials were given advance information by management of the impending establishment of new or additional positions which would soon have to be filled. In general, it was expected that personnel officials would draw upon both types of data sources for identifying total organizational staffing needs.

Increasingly, however, these traditional methods have proved insufficient. As the Federal workforce has continued to evolve from a largely clerical structure to a largely technical and professional structure, the average Federal employee skill level--and thus the average qualification requirements level--has steadily increased. This has meant a corresponding increase in the lead time necessary to train the average Federal hire up to full-performance level. Lead times of 2-3 years for development of an average full-performance worker are now the rule, not the exception. This means that in many fields, trainee hiring must increasingly be keyed, not to current vacancies, but to full-performance vacancies that will be coming up 2-3 years from now.

Another recent development has added greatly to the need for anticipating future staffing needs: the advent of multi-year Federal program planning. Since so much of an organization's personnel management program is a direct function of the organization's staffing needs, the effective multi-year planning of such personnel programs clearly requires that future changes in staffing needs be identified well in advance of their occurrence--a task for which the traditional methods of staffing needs identification are clearly inadequate.

For a variety of reasons, then, contemporary conditions require that Federal personnel managers develop methods for anticipating future organizational staffing needs and that these methods be made a part of the personnel management function at all levels.

This anticipation of future staffing needs--which is called staffing needs planning--is one part of the overall process by which an organization plans for its manpower. This total process is called organization manpower planning. To get a proper perspective on the place of staffing needs planning in the total system, we must carefully define some terms which have in the past been used imprecisely.

Definitions and Responsibilities

The term "manpower" combines the element "-power", or capacity to do work, with a collective element indicating the source of that power: i.e., "man-", in its usual generic sense of "human beings." Originally used to refer to the aggregate work capabilities of the total labor force of a state or nation, "manpower" has come to be applied to the collective labor force of any identifiable functional entity--an organization, industry, company, etc. Adding to this the term "planning" gives us "manpower planning", a term in common use throughout the English-speaking world (and in literal translation, throughout the Westernized world) for "the systematic planning of and for the manpower requirements of a specified functional entity."

Over the years, "manpower planning" has been applied to a variety of rather different types of activities. One type is oriented toward labor force analysis. This includes both (a) the development of descriptive statistics showing the significant dimensions and components of a specified labor force and their change over time, and (b) the conducting of analytical and projective studies of the changes in, and the factors relating to, the labor force features thus portrayed. This type of activity is typified by the work of the Bureau of Labor Statistics and the Bureau of the Census.

A second type of activity is oriented toward the development, administration and evaluation of manpower programs. This includes both (a) programs to improve the employment status of particular groups or segments of the labor force (e.g., equal employment opportunity programs), and (b) programs to promote the optimization of manpower supply/demand relationships in particular industries and/or occupational areas.

Neither the analysis-oriented nor the program-oriented types of manpower planning, however, have much in common with organization manpower planning: the systematic planning of the manpower needs of individual organizations. Organizational manpower planning has a different purpose, uses different types of data, and--especially--uses different specialized methods and techniques than does either of the other major types of manpower planning.

Organization manpower planning is made up of two quite distinct aspects.

The first aspect of organization manpower planning is the planning of the numbers and kinds of workers needed to perform the organization's work. This is workforce planning.

The second is the systematic estimation of the numbers and kinds of future personnel management actions which will have to be taken in order to provide this required workforce. This is staffing needs planning.

In theory, these two functions interact to form a continuous, coherent organization manpower planning process. In practice, however, these two functions operate, in most government organizations, relatively independently.

Workforce planning--deciding what types of workers, and how many, are to be present in the organization--is a responsibility of agency management. The manager may have help in performing this function from such management staff as "O&M" (organization and methods), "OR" (operations research), or budget specialists. And in many agencies management officials, as a matter of policy, consult personnel officials on the manpower aspects of workforce planning. (Cf., pp 8-9, below.)

Staffing needs planning, on the other hand--planning the future personnel managment actions needed to provide the manager's required workforce--is a responsibility of the agency's personnel director.

These two functions also differ in their purpose and methodology. The purpose of workforce planning, for example, is to answer the questions:

-- What kinds of workers will be needed?

-- At what skill levels? And

-- How many of each?

The purpose of staffing needs planning, on the other hand, is to answer the questions:

-- What types of future personnel actions, and how many, will be needed to provide management's planned workforce?

-- Will providing the workforce be feasible? (If not, what changes will be needed?) And

-- What will providing it cost?

On the subject of methodology, we would point out that the methodology of workforce planning, in its most typical form, is to start with (a) estimates of expected workload, to which are applied, (b) measures or assumptions of output per unit of labor time, and by this means to derive estimates of (c) the numbers and types of workers needed to produce the expected workload. Workforce planning methodology, that is, uses workload and work measurement data and derives "required workforce" plans.

The methodology of staffing needs planning, on the other hand, is to start with (a) the manager's "required workforce" plan, estimate (b) the personnel losses and shifts likely to take place during the planning period, and then to determine (c) what future personnel management actions will be needed to provide the required workforce. Thus, staffing needs planning uses mostly personnel data and produces either suggested changes in management's workforce plan or summary estimates of what must be done to provide the required workforce.

It is the purpose of this handbook to set forth in substantially full detail (a) the policies and features which characterize effective staffing needs planning programs, and (b) a specific basic technology for performing the key analytical functions of such programs.

Suggested Policy Requirements

From the above discussion of the distinction between workforce planning and staffing needs planning, it is clear that a definitive manpower planning policy is needed which will make explicit both (a) how these two functions differ, and also (b) in what respects they are interdependent and must function in close coordination.

In general, we suggest that an effective overall policy would provide that:

1. Personnel management officials should recognize the workforce planning aspect of organization manpower planning as the direct responsibility of agency management; that
2. Such personnel officials should provide information and assistance to management and management staff in the performance of their workforce planning responsibilities under policies and requirements established by higher management authority, the Office of Management and Budget, the President and the Congress; that
3. Agency management officials and management staff should recognize the staffing needs planning phase of manpower planning as the direct responsibility of organization personnel management; and that

4. Such management officials should provide information and assistance to personnel officials in their performance of the staffing needs planning responsibilities under policies and requirements established by higher authority, the Civil Service Commission, the President and the Congress.

In addition to a general policy, guidelines for the basic functions of staffing needs planning programs should be spelled out in concrete detail. We suggest that specific reference be made to the following:

1. Information Functions - There should be provision for:

   - The regular transmission to organization personnel officials of detailed data on the workforce structure established or proposed by management for each phase of the organization's planning period; and

   - The regular transmission to organization management officials of the types of personnel management and staffing needs planning information detailed under "Workforce Planning Functions," below.

2. Analysis Functions - Provision should be made for the regular, scheduled analysis of manpower resources and personnel transactions data. This may include such specific analyses as:

   - Trends in the occupational, grade or skill level, and/or pay distribution of organization workers in specified occupations, functions and/or organization segments;

   - Levels and trends of workforce composition in particular occupations or functions by such dimensions as age, sex, minority status and/or length of service;

   - Levels and trends of employee retirement losses and/or eligibility;

   - The pattern and/or trend of workforce accession actions (number and percent of outside hires at entry levels, etc.);

   - Levels and trends in occupational advancement patterns, inter-occupational flows, etc; and/or

- Levels and trends of occupational and/or organizational loss rates due to quits, transfers, occupational shifts, etc.

3. Staffing Needs Estimating Functions - Provision should be made for the systematic estimation of current and future staffing needs in key organizations, occupations or specialties. This may include such specific activities as:

   - Analyzing the net workforce changes which will be required during each phase of the planning period under management's actual or proposed workforce plans;

   - Projecting the numbers and types of vacancies that are likely to occur during each phase of the planning period due to such causes as turnover losses, death, disability, retirement, management actions, etc.;

   - Projecting, by and/or with the aid of personnel functional specialists (staffing, training, etc.), the numbers and kinds of personnel management actions which will be necessary under current or proposed personnel policies to provide the required workforce when, where and with the skills needed; and

   - The preparation in convenient, hard-copy form of written summaries of the results of such staffing needs estimating activities.

4. Workforce Planning Functions - Provision should be made for the development and the transmission to management of personnel management and staffing needs data and analyses which are needed for the effective performance of workforce planning functions. This may include such data and analyses as:

   - Detailed assessments of the feasibility of providing management's proposed workforce, based on labor market limitations, the numbers of employees in the training pipeline, etc.;

   - Detailed analyses of the means necessary to staff the proposed workforce (extent of outside hiring at entry level necessary, amount of employee training or re-training needed, etc.);

- Estimates of the direct and indirect costs of the necessary personnel staffing actions (cost of recruitment, training, employee relocation, separation and leave payments, etc.);

- Information on the impact on workforce cost estimates of new employee salary or pay schedules, job or occupation reclassification actions, employee grade distributions, etc.;

- Analyses of the impact of the proposed program of staffing actions on (a) the existing workforce (advancement or retention opportunities for women, minorities, handicapped, etc.) and on (b) the organization's responsibilities for implementing public policies (EEO, upward mobility, older workers, veterans, etc.); and

- Recommendations for changes in (a) management's workforce plans (occupational or skill-level tradeoffs, etc.) and/or in (b) organization personnel policies or practices (policy on outside hiring, etc.), based on the above.

5. Personnel Program Planning Functions - Provision should be made for the utilization of the above-described data and analyses in the establishment of operational goals and objectives for the various personnel management functional activities and in the development of personnel management budget estimates.

6. Data System Functions - Provision should be made for an effective data system for obtaining, recording and furnishing the data needed to support staffing needs planning functions. Such systems should include specific provisions for obtaining and recording needed data on the workforce plans and proposals made by organization management as part of the workforce planning and budgeting process.

7. Evaluation Functions - Staffing needs planning systems should make specific provisions for the regular and systematic evaluation of the above-described manpower planning policies, procedures and products, and of their contribution to the overall planning, budgeting and personnel management systems of the organization.

Plan for This Handbook

As will be appreciated, addressing all of the above functions in one single basic handbook is not feasible. Accordingly, this handbook focusses on what is perhaps the pivotal technical function of staffing needs planning: the estimation of future vacancies and, in particular, on the analysis and projection of employee turnover.

For more than seventy years American researchers have sought in vain for a simple, reliable method for analyzing and projecting turnover. Scores of methods have been tried. But each has failed when tried in new situations--or even in the same situation after a lapse of time. Now, however, effective methods are available.

The methods set forth in this handbook had their origins in basic research done at the U.S. Civil Service Commission during the period 1966-70. [1] This research has been subsequently confirmed as being consistent with long-term actuarial studies in Great Britain and on the Continent and the methods can be considered proven effective and reliable.

Needless to say, the authors of this handbook accept full responsibility for the technical effectiveness and adequacy of the specific methods described. Any errors or weaknesses are ours alone.

In preparing the various sections of this handbook, we have tried to follow a four-step procedure. First, to describe briefly the specific nature of the problem at hand. Second, to describe more or less in detail the technology which applies to the problem. Third, to provide a detailed, step-by-step method for making the necessary calculations or manipulations. And fourth, to describe the features and capabilities of the computer programs provided to perform each analytical step.

A number of considerations were involved in the design of the computer programs. From the outset, we determined that they should be:

1. General in application -- usable by anyone in any organization;

2. Complete and self-standing -- all calculations, including statistical tests, to be done by the program; no statistical background required to operate them with full effectiveness;

---

[1] Reported to a September 1971 NATO Conference in: H. L. Clark, "Problems and Progress in Civil Service Manpower Planning in the United States," reprinted in Manpower Planning Models, Clough, Lewis, and Oliver, Eds., English Universities Press, London, 1974.

3. Technically compact and simple — programs are written in Fortran IV, one of the most common computer languages, and use only elementary commands and a minimum of core; any time-sharing service can run them; and

4. Fully-documented -- For each program, we provide:

   a) Program Listing--a complete listing of the entire program and all required subroutines.

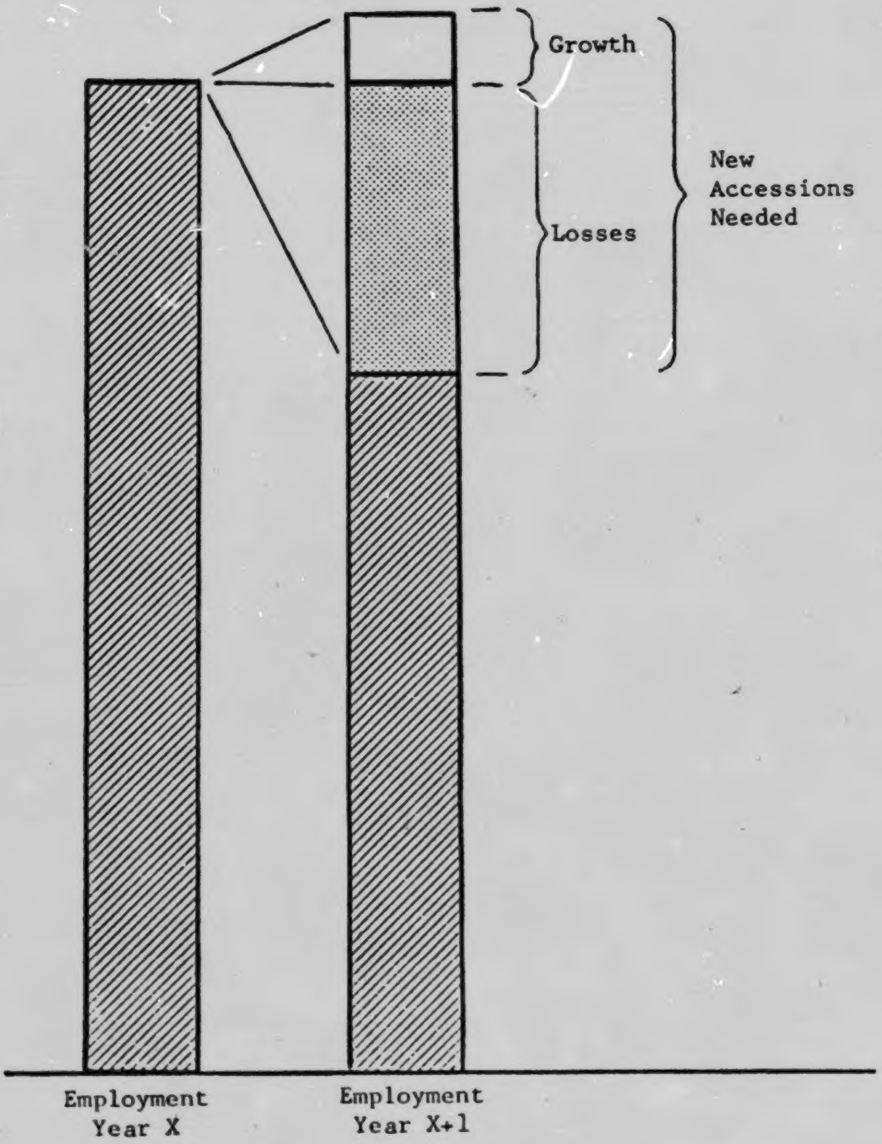
   b) Operation Manual--a complete set of step-by-step instructions for the program operator.

   c) Technical Analysis--a detailed analysis of what the program is doing and how results are obtained.

Taken together, we believe that this handbook will give both the newcomer and the experienced worker a sound grounding in the central analytical tasks of staffing needs planning. Prior training in statistics is not assumed or required--though, needless to say, it will be helpful if the reader has it. Some aptitude for quantitative work, as well as some elementary algebra, are required.

For all, we hope this handbook will open new doors to solutions to old problems. If it does, we shall feel amply repaid.

Figure 2-1


Growth
Losses
New Accessions Needed
Employment Year X
Employment Year X+1

CHAPTER 2

ANALYZING WORKGROUP LOSSES

The Technical Problem

The pivotal technical problem in the estimation of future vacancies is the analysis and projection of turnover losses. This can be clearly seen from Figure 2-1.

Notice that the group of positions labelled "Growth" can be readily determined by simple subtraction of the starting population (left bar) from the projection-period's ending population (right bar). What the level of "Losses" will be, however, must be estimated by some other means.

Means for estimating some kinds of losses are readily available. Losses from death, disability, and retirement are termed actuarial-type losses because actuarial tables for estimating such losses have been available for many years. These are simple "annual loss probability" tables based on the employee's age and sex. Samples of such tables will be provided later. Here, it is only necessary to note that tables for making such estimates are readily available.

There are no actuarial tables available, however, for other types of personnel losses (voluntary quits, etc.). A variety of techniques have been tried over the years to deal with such losses. But it has only been recently, with the development of automated personnel data files, that data have been made available to permit effective large-scale research to be conducted.

The U.S. Civil Service Commission has been actively engaged in manpower planning research, and particularly in the study of turnover, since the establishment of the Federal Personnel Statistics File (The "10% Sample") in 1962. This File contained a continuous work history sample of all Federal employees whose Social Security number ended in "5." From this initial 10%, this sample has been expanded to its present 100%. In both its 10% and 100% forms, this sample has made it possible to study turnover in a degree of detail never before possible.

It had long been known, of course, that turnover rates differed substantially by occupation. With the advent of a computerized data base, however, we have been able to study turnover in particular occupations in great detail. And this research has taught us a great deal.

Before we take up what has been learned from this research, one preliminary point should be clearly understood. In this discussion--and indeed throughout all portions of this handbook--we are dealing with retention and turnover of non-temporary employees only. Losses among temporary employees, such as separations at the end of the job's planned duration (e.g., termination at the end of an appointment "not to exceed 90 days") are not dealt with in this pamphlet.

With it clearly in mind that we are talking about continuing employees only, then, let us see some of the things that have been learned from turnover research.

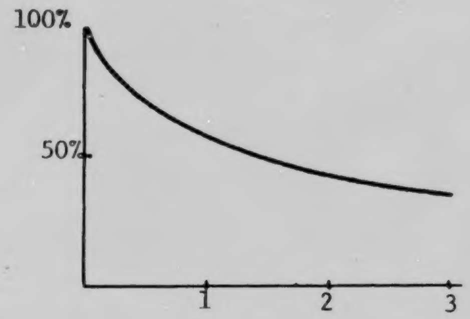
The Results of Turnover Research

Just about the first things you learn when you begin to study occupational turnover[1] closely are that turnover is a function of employees' lengths of service, and that most turnover occurs in the period immediately after hire. Additionally, you learn that the turnover curve is a very peculiar kind of animal indeed.

For example, some years ago we took a group of low-level clerical hires and followed them over a three-year span, counting at the end of each year how many were left of the original starting group. The figures we got looked something like this:

Start	100%
End of 1st year	58%
End of 2nd year	44%
End of 3rd year	36%

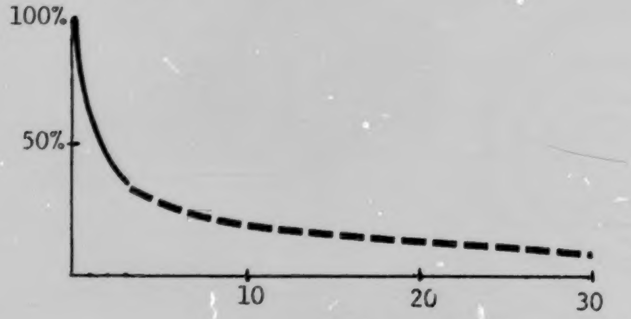
When graphed, these points look like this:



---

[1] The occupational analyses described in this handbook do not need to be done for every occupation within an organization. They may be restricted to an organization's major occupations. Statistically, the techniques are more effective for occupations (or groups of hires) with 32 or more employees. By use of the techniques in Chapter 8, it is possible to combine smaller occupations into larger groups if a user so desires.

Now, we know from long actuarial experience based on our Civil Service Retirement Fund records that about 5% of such hires remain in the service all the way to retirement age. Thus a full 30-year curve for this group must look something like this:



That's a rather striking looking curve, isn't it?

Let us look at some of its properties. First, we see that 64% (100% - 36%) of the total of 95% who will leave before retirement are already gone by the end of the third year after original hire. And of these, 42%--or more than half of the 3-year total losses of 64%--left in the first year alone.

When we first started seeing curves like this, we made up a simple "rule of thumb" to describe this concentration of losses in the early years. This "rule of thumb" went something like this:

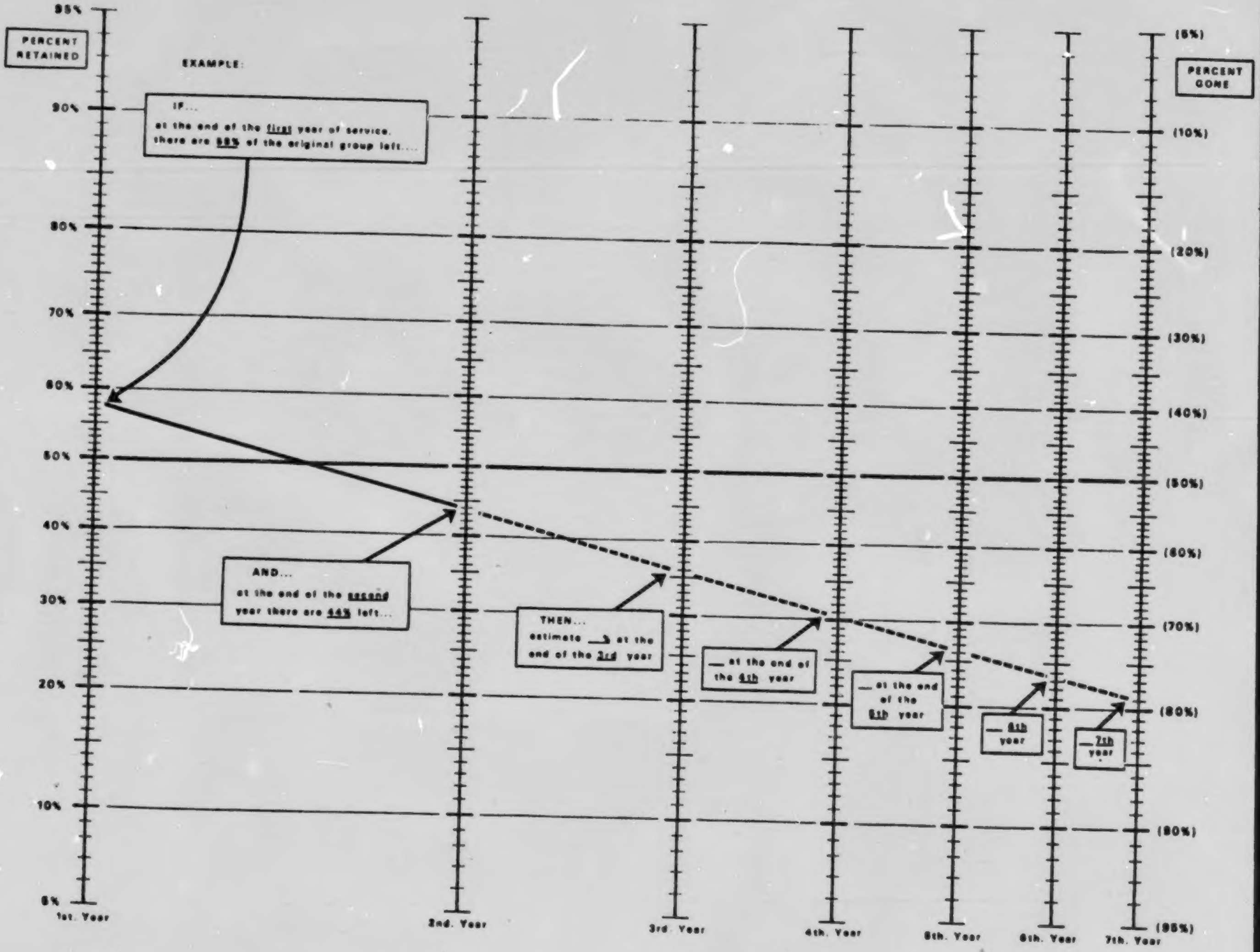
> "In any given group of hires, from two-thirds to three-fourths of all of the quits that will ever occur prior to retirement age will have already occurred by the end of the first three years of service. And of these, more than half will have occurred by the end of the first year alone."

We found this rule of thumb useful in getting across to people what a high proportion of quits occur in the first few years of service. And it was useful in showing that the data needed to measure turnover were retention rates over time--i.e., longitudinal data. But it really was not much help in a computational sense.

The second major thing we can see from this type of data is that the turnover curve is not the simple kind of curve people have thought it was. For example, in this sample we see that 42 out of every 100 in our starting group left in the first year, another 14 left in the second year, and 8 more left in the third year. Obviously, then, we cannot just say that any one fixed number of people are leaving our group during each year of service.

BUREAU OF POLICIES AND STANDARDS, U.S. CIVIL SERVICE COMMISSION

LOG-PROBABILITY NOMOGRAPH FOR PROJECTING WORKGROUP RETENTION (AND LOSS)



EXAMPLE:

IF...
at the end of the first year of service, there are 58% of the original group left....

AND...
at the end of the second year there are 44% left...

THEN...
estimate __% at the end of the 3rd year

__ at the end of the 4th year

__ at the end of the 5th year

__ 6th year

__ 7th year

Well, how about a fixed percentage of people leaving each year, say, "X% of the people on board at the start of the year will be lost during the year"? This is the kind of thing most people think of when they simply take the overall annual loss rate of the workforce as an adequate measure of "the turnover rate." This is also the approach used in some kinds of statistical modelling techniques utilizing transition matrices or "Markov Chain" methods.

The retention data show that this approach will not work either. In the first year, for example, we lose 42 out of 100 or 42%. In the second year, 14 out of 58 or 24%. And the third year, 8 out of 44 or 18%.

What we can see from these data, then, is that (a) turnover is by far at its heaviest in the first three years after hire--and especially in the first year--but (b) the shape of the turnover curve is by no means the simple type of thing most people think it is.

The Log-Probability Nomograph

To show you just what the turnover curve really is, we must look at what is called the "log-probability nomograph" (Figure 2-2). The instructions for using it are written right on it.

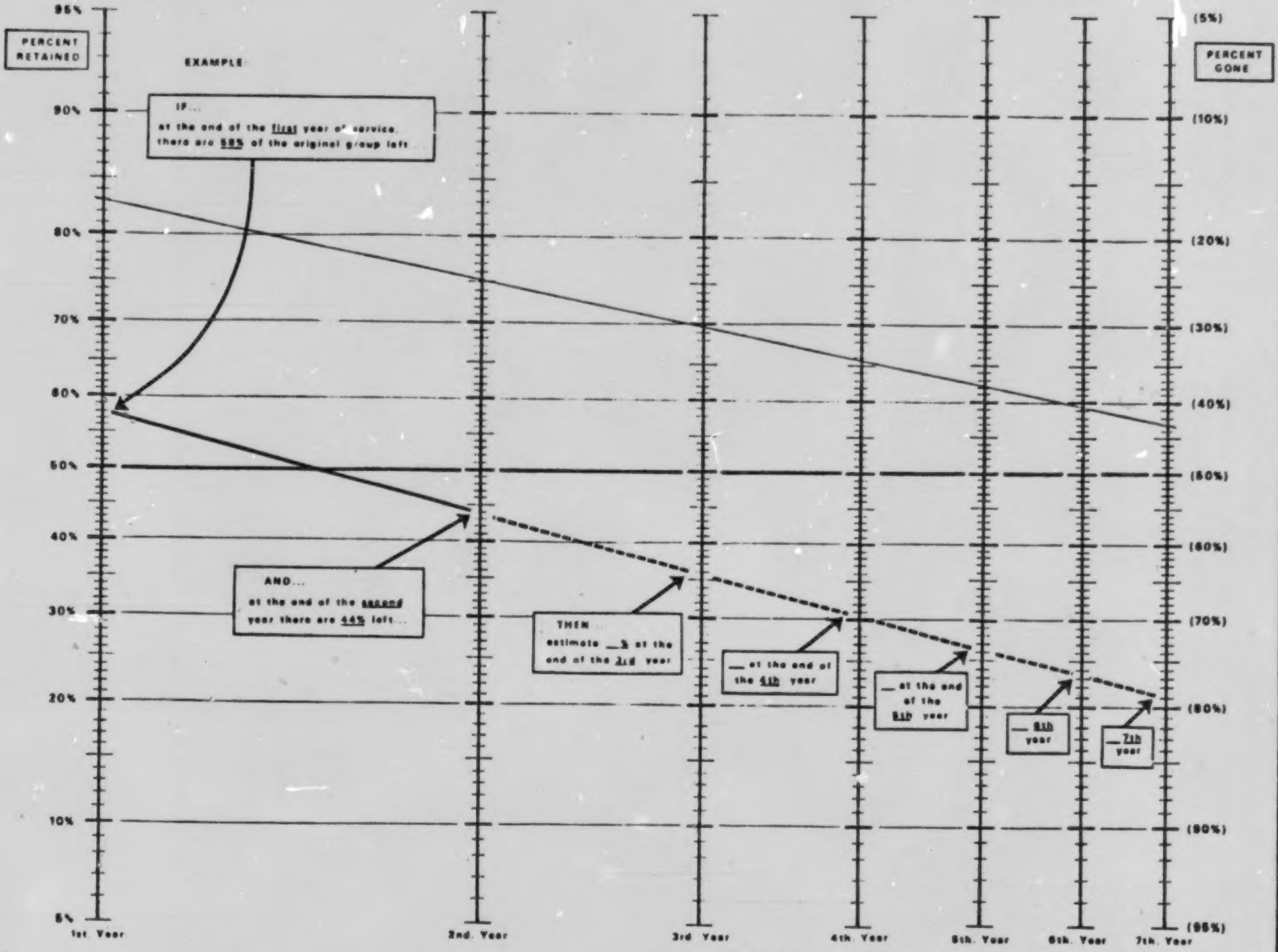
> First, you obtain data on what percent of a given group of hires are still on board one year after the date of original hire and locate this point on the "1st year" percentage scale.
>
> Second, you take the percent of the original group that are left after the second year and locate this point on the "2nd year" percentage scale.
>
> Then you take a straightedge and draw a line through these points and on through the remaining years' scales. Where this line hits the scales for the 3rd through 7th years indicates the percentage of the original starting group that are likely to be left after each year.

As you see, a sample line has been drawn on the nomograph using the data points that we gave earlier. From this sample line, we can now extend the number of data points we have up to seven:

Year	% Left
1st	58
2nd	44
3rd	36
4th	30
5th	26
6th	23
7th	21

BUREAU OF POLICIES AND STANDARDS, U.S. CIVIL SERVICE COMMISSION

LOG-PROBABILITY NOMOGRAPH FOR PROJECTING WORKGROUP RETENTION (AND LOSS)



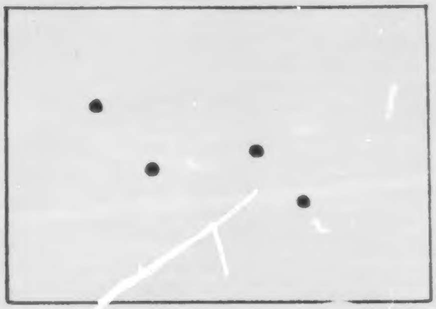
Having seen one turnover curve for one group of employees, let us plot the curve of another group. This time we take a group of professional hires and we find that in this group some 83% are left at the end of the first year and 75% are left at the end of the second year. Plotting these two points on the first two scales and drawing our straight line, we read from the remaining scales (Figure 2-3):

Year	% Left
3rd	70
4th	66
5th	62
6th	59
7th	57

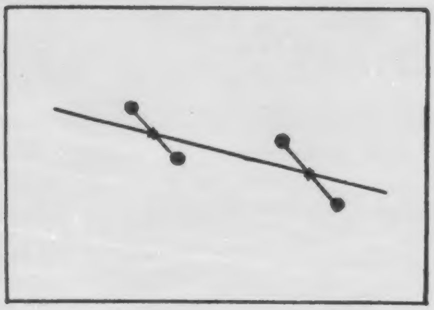
From these two examples it is clear that the log-probability nomograph can be used to plot the long-term retention curve for any cohort (i.e., starting group) for which we have longitudinal data (i.e., measurements over time).

Suppose, however, that we have more than just 2 years' data--say, three or four years--and that when we plot them we find that they are not quite in a perfect straight line. (Since there is almost always some random variation around any norm, this can be expected to be the usual case.) When this occurs, all of the available points should be used in drawing the final straight line.

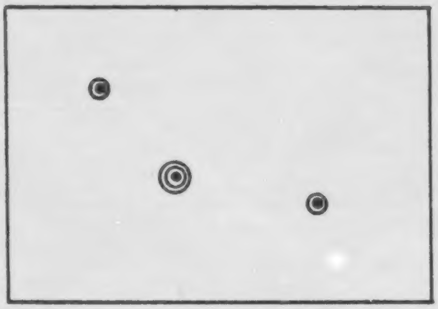
Suppose, for example, that we have four points like this (the vertical separation of the points is exaggerated here for the sake of clearer illustration):



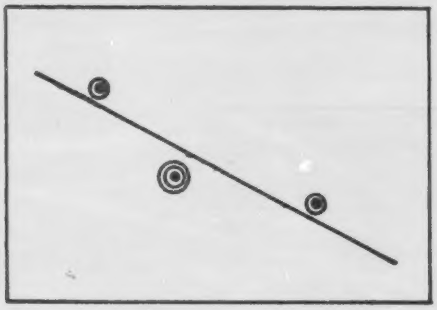
In this case, we simply reduce the number of points to two by connecting the first two points together and the last two points together with straight lines and locating the midpoint of each line. The final line is then drawn through these two midpoints, thus:



If there are only three points, the solution is not quite as satisfactory, but will give a reasonably good approximation of the correct line. For three points, we take a pencil or other suitable marker and draw one small circle around the first and last points and two circles around the middle point, like this:



Ideally, you should then be able to draw the final line tangent to the circles around all three points. If they are still too separated for that, however, the final line should be so drawn that the total distance from the line to the circles around the first and third points combined is as nearly as possible equal to the distance between the line and the outer circle of the middle point alone, thus:



For more than four points, there really is no simple method other than to say that the sum of the distances from the line to points above the line should be as nearly as possible equal to the sum of the distances from the line to the points below the line. Obviously, the nomograph method is not well suited to large numbers of data points.

Application to Hiring Levels

Let us return to the data table we developed using the first set of retention data. There are a great many useful points that this table can illustrate.

First, let us apply this turnover curve to the case of an organization that has been hiring exactly 100 employees in this occupation each year for the last 6 years. A length-of-service distribution of the employees still on board would look like this:

	Years Since Hire	Number
	6	23
	5	26
	4	30
	3	36
	2	44
	1	58
Total		217

As you see, this looks just like our original retention table and, if we graph these values, the result would look just like our turnover curve. So the first thing we can learn from studying the turnover curve is that--other things being equal, of course--the length-of-service distribution of an occupation's workforce will tend to look like that occupation's turnover curve.

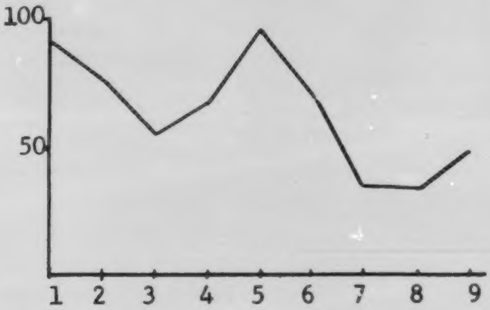
In actual practice, of course, most organizations don't tend to hire the same number of people every year. Rather, you hire more some years and less in others. But though these ups and downs in hiring cause corresponding ups and downs in the length-of-service distribution, the turnover-type curve can usually be found if you know how to look for it.

For example, one organization's length-of-service distribution for employees who had completed at least one year of service looked like this:

Years	Number	Years	Number
1	92	6	70
2	77	7	35
3	55	8	34
4	67	9	48
5	95		

(We will use only a few years of data to keep this simple.)

Obviously, when we plot these points they do not look very smooth:



The ups and downs can be smoothed out, however, by using a quadratic centered-moving-average technique. To do this, you choose an odd number of points--say, three or five--and taking the middle point as the center, find the quadratic mean of the center point plus either one or two points on either side. In this case, for example, we use a 5-point centered-moving-average and we compute the first points like this:

CMA for year	is equal to
3	$\sqrt[5]{92 \times 77 \times 55 \times 67 \times 95}$
4	$\sqrt[5]{77 \times 55 \times 67 \times 95 \times 70}$
etc.	etc.

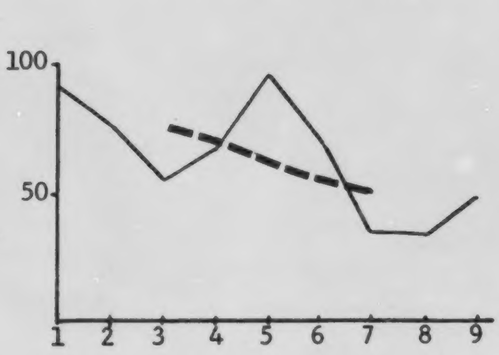
This is most easily done on a calculator:

Step 1. $92 \times 77 \times 55 \times 67 \times 95 = 2.47993 \times 10^9$
Step 2. $\log (2.47993 \times 10^9) = 9.39444$
Step 3. $9.39444/5 = 1.878888$
Step 4. antilog $1.878888 = 75.7$

And so on. By this means we find that using a 5-point quadratic c.m.a. for these points gives:

Center Pt.	C.M.A.
3	75.7
4	71.6
5	61.2
6	55.6
7	52.0

Plotting these points on our original graph gives:



As you can see, the smoothed data points look much more like the turnover-type curve we were looking for. As you can also see, however, the tendency in this organization is for new hiring to go sharply up and down in quite regular cycles over a multi-year period. Many organizations show such cycles and their identification, through study of length-of-service distributions, can often be helpful in projecting future hiring ups and downs.

Let us return, however, to our hypothetical length-of-service table and see what we can say about the level of turnover we can expect to occur in this group in the next year. Using our turnover curve, we estimate the following losses:

Years Since Hire	Number	Number After 1 Year	Lost In Year
6	23	21	2
5	26	23	3
4	30	26	4
3	36	30	6
2	44	36	8
1	58	44	14
	217	180	37

Thus, we estimate turnover over the next year as 37 out of 217 or 17.1%.

Now, however, there is a very important point we can show. Let us illustrate it this way. Suppose we decide that we are going to go out and hire another 100 people now. What effect will this have on next year's turnover? Let us add these new hires to our table as follows:

Years Since Hire	Start of Year	No. After 1 Year	Lost in Year
6 - 1	217	180	37
0	100	58	42
Totals	317	238	79

As you see, our turnover estimate for next year is now 79 out of 317 or 24.9%.

Suppose we tried other levels of new hires such as:

Number of Hires	Expected Losses		
	Number	Out Of	Percent
0	37	217	17.1
50	58	267	21.7
100	79	317	24.9
200	121	417	29.0

The important point these data illustrate--and it is a very important point indeed--is that the turnover rate of any given group is directly dependent on the rate of new hires being added to the group. If new hiring goes up or down, so does the group's overall turnover rate.

One reason this point is important is that it gives us an explanation for something that labor economists have often observed: The tendency for government's overall turnover rate to go up when the economy rises and down when the economy declines.

What happens is not that there is any change in any existing group's length-of-service loss curve--on the contrary, these curves tend to hold their steady course no matter what the state of the economy. Rather, what happens is that what goes up in good times and down in bad times is the overall level of government hiring.

When you hire more, you have more new recruits in your workforce and, since these are lost at much higher rates than your longer-service workers, your overall loss rate goes up. When you hire fewer new people, the process is exactly the reverse. Fewer new recruits mean fewer turnover losses and a lower overall turnover rate.

A second reason this point is important is that it gives us new insight into what causes differences or changes in the turnover rates of particular organizations or groups. Very often people interpret changes or differences in turnover rate as reflecting poor morale, outside competition, inadequate pay rates, bad management, poor organization practices, etc., when almost invariably the real causes lie in hiring practice differences or changes.

And finally, this point is important because it shows that the nomograph-type of turnover projection can predict an organization's future turnover levels even when the number of new hires to be added to the workforce (or, the total size of the workforce) represents a major departure from the organization's past trend.

This is a virtue of very special usefulness. It is one thing, after all, to be able to predict future turnover needs when the organization is on a stable trend, with little change from year to year. Any trend-projection technique from "rules of thumb" to Markov chains can make good projections in these circumstances. But it is quite another matter to predict turnover when abrupt and unprecedented change from past trends is in the works. Yet it is at times of such abrupt change that effective planning is most needed. Techniques that can handle such change situations, therefore, are very useful indeed.

Other Applications

Although we have now seen what is perhaps the most important conclusion to be gained from discussing the nomograph, we should not conclude without at least a brief mention of some of the other useful applications that can be made of our turnover curve table.

First let us return to the basic table we made up for the turnover curve shown on the nomograph:

Year	% Left
Start	100%
1st	58
2nd	44
3rd	36
4th	30
5th	26
6th	23
7th	21

From this basic table, we can compute:

1. The probability of a new hire's lasting x years. E.g., the probability of a new hire staying 6 years is 23/100 or 23%.

2. The probability of an employee of x years service staying until year y. E.g., the probability of an employee with 3 years staying to the end of 6 years is 23/36 or 64%.

3. How many people have to be hired now to have x people on board y years from now? E.g., the number of hires needed to have 50 people on board 3 years from now is equal to:

33

$$\frac{x}{50} = \frac{100}{36}$$

$$36x = 5000$$

$$x = 5000/36 = 139$$

4. Assuming a training course costs \$1000 per employee, how much money must be spent on training new hires now for every employee needed on board 3 years from now? E.g., the cost of the training divided by the 3-year retention probability is equal to:

$$\$1000/.36 = \$2778$$

As you see, there are a number of different kinds of computations that can successfully be made from just the limited amount of data in our basic table.

Conclusions

Let us conclude our discussion of the nomograph method by summing up this method's advantages and drawbacks.

On the one hand, it is easy, quick, and cheap. The nomograph shows right on it the data that are needed and how to draw turnover curves. Anyone can use it--it requires no mathematical or statistical expertise. Anyone who has the ability to read a scale and draw straight lines can make seven year projections from simple data.

These characteristics make it very handy to use in small groups, under field conditions, or by such skilled scale-readers and draftsmen as crafts-and-trades employees or foremen. Also, since the seven scales can be taken as representing any equally-spaced time units--seven weeks, seven months, seven quarters, etc.--the nomograph method can be used on shorter-term, as well as longer-term problems.

Finally, the method uses data that usually are readily available in organization records. Even in organizations without detailed personnel data systems, there are usually pay records showing the number and types of employees hired in a given period and how many were still on the rolls at selected subsequent time periods.

These advantages make the nomograph method very useful in many common types of situations. It is sufficient, indeed, for many practical situations.

The nomograph method also has, however, some truly significant drawbacks:

- Manual line-fitting gets very hard if you have more than four points;

- When retention observations are unevenly spaced, the correct placement of points in between scales is hard to achieve;

- Nomograph results are difficult to apply directly where the group's hires have been made over a span of time rather than being bunched together;

- The nomograph curve directly applies to employees in the first seven years of service only;

- It is hard to apply nomograph curves to mixed length-of-service groups, such as an organization's overall workforce with its typical mixture of all different lengths of service; and

- While the nomograph method shows the most probable level of future retention, it is difficult to estimate how reliable that projection is and how much variation can be expected from these projected values.

Clearly, while the nomograph method is both useful and instructive, effective staffing needs planning requires more sophisticated, more flexible, and more powerful methods. Such methods have been developed and are available. And it is to them that we now turn.

CHAPTER 3

ON CURVE FITTING

During our discussion of the log-probability nomograph, you saw how approximation methods could be used to fit a straight line to a plotted set of retention data. The line we were looking for there was the one which would best represent all of the points. The purpose of the statistical method described in this chapter is exactly the same as that used on the nomograph--finding the best line.[1]

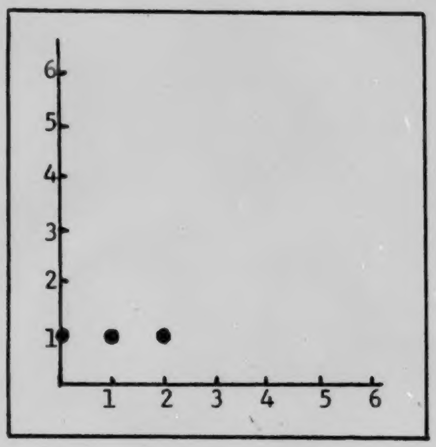
To develop this method, we need to know how to:

1. Determine the formula for a straight line from a given set of data points; and
2. Transform longitudinal retention data so that it can be fitted with a straight line.

The General Formula For A Straight Line

To begin the discussion of a statistical method of fitting a straight line to a given set of data points, we must shift our attention from points plotted on a nomograph to points plotted on a standard graph (such as the one on which our original retention curve was plotted). This graph consists of two perpendicular lines. The vertical line is called the y-axis; the horizontal line is called the x-axis.

Consider first the following graph:



---

[1] Readers familiar with the linear least-squares curve-fitting method, described in this chapter, may proceed to Chapter 4.

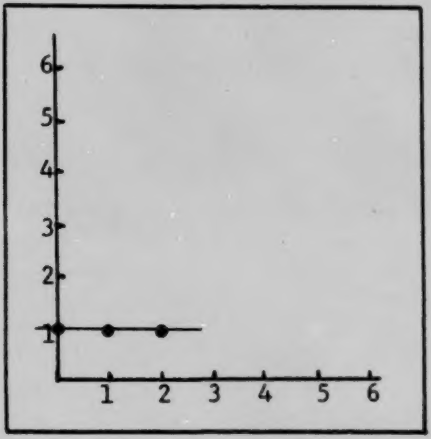
38

The three points plotted on the graph can be represented by the following table:

x-axis	y-axis
0	1
1	1
2	1

This means, for example, that the value 0 on the x-axis is matched with the value 1 on the y-axis. We will refer to the values on the x-axis as x-values and to those on the y-axis as y-values.

Connecting the three points on the graph gives a line which is parallel to the x-axis:

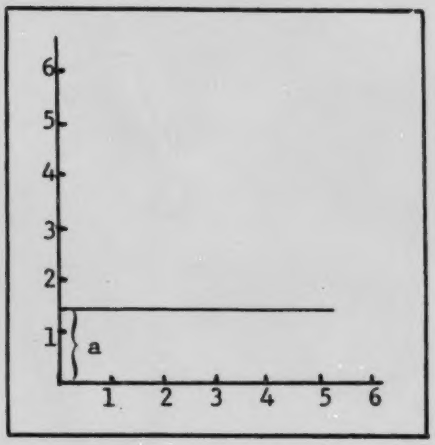


Since for each value on the x-axis (i.e., for each x-value) the corresponding y-value on the line is 1, the formula (or equation) describing this line is written:

$$y = 1$$

Note that this relation also holds true for $x = 0$. It can be said that this line intercepts (or hits) the y-axis at 1.

To generalize this concept, consider any line parallel to the x-axis and let the distance from the x-axis to the point where the line intercepts the y-axis be represented by the letter a as follows:



Using the same reasoning as before, we can say that this line is represented by the equation:

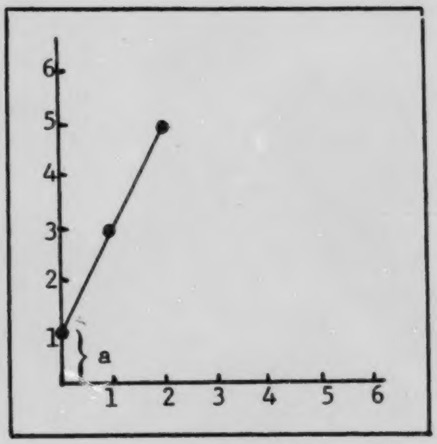
$$y = a$$

The point at which a line intercepts the y-axis is called, quite naturally, its y-intercept.

Now suppose we make a slight change in our set of points and use:

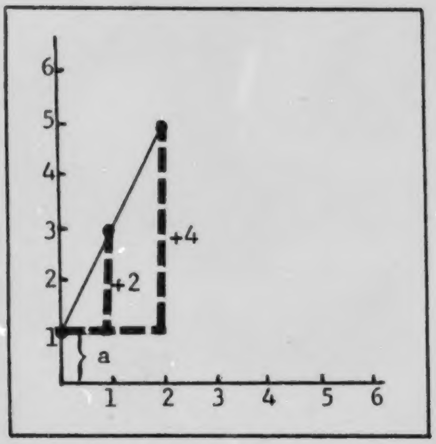
x	y
0	1
1	3
2	5

Then our line looks like this:



We still have a value for a (the y-intercept) of 1 but obviously something else is needed. This "something else" is a factor which measures the line's departure from being parallel to the x-axis. Or, in other words, a factor which measures the amount by which y differs from a.

Consider the following diagram:



Note that for x = 1, y = 3, or y = 1 + 2= a + 2. And for x = 2, y = 5 = 1 + 4 = a + 4. These two values of y can also be written still another way:

$$y = a + 2 = a + 2(1)$$

$$y = a + 4 = a + 2(2)$$

Note that the values in the parentheses are the corresponding values of x.

So, for x = 1,

$$y = 3 = 1 + 2 = a + 2(1) = a + 2x$$

And for x = 2,

$$y = 5 = 1 + 4 = a + 2(2) = a + 2x$$

Now, replacing the number 2 by the letter b, we have:

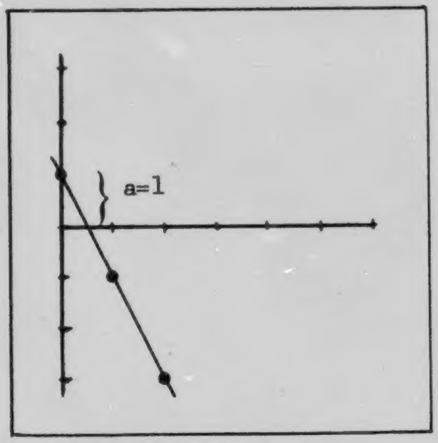
$$y = a + bx$$

This is a generalized equation for a straight line. And b is referred to as the slope of the line.

Our example above showed b as a positive number. But b can also be negative. For example, consider the points:

x	y
0	1
1	-1
2	-3

The graph of these points is:



The value of a is still 1. But in this case, for x = 1,

$$y = -1 = 1 - 2 = 1 - 2(1)$$

And for x = 2,

$$y = -3 = 1 - 4 = 1 - 2(2)$$

So in this example the equation is:

$$y = 1 - 2x$$

And b = -2. (In our analysis of retention data we will find that the value of b is always negative.)

To sum up then, we have:

A general formula for a straight line is:

$$y = a + bx$$

where

a = the y-intercept
b = the slope

Once the equation for a line is known, it is possible to use that equation to solve for a value of y for each value of x.

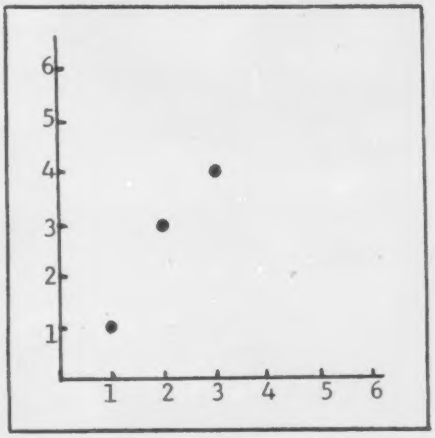
The Linear Least Squares Line

Now that we have a general formula for a straight line, we want to be able to use it to find the line which best fits a given set of data points. To do this we need a method to calculate values for a and b using the x- and y-values of the data points.

As an example, let's use the points:

x	y
1	1
2	3
3	4

You can see from the graph below that these points are not exactly in line:



The first step in the procedure to find the best line to fit these points is to set up a work table with these headings:

x y xy x^2

The first column of the table will list all of the x-values. The second column will list the y-values (with each given x, y pair--i.e., point--on the same line of the table). For the third column, each x-value is multiplied by its corresponding y-value and for the fourth column, each x-value is multiplied by itself (i.e., squared). After the columns are filled in, the values in each one are added to obtain a total figure for each column.

Using our sample set of points, a work table would look like this:

x	y	xy	x^2
1	1	1	1
2	3	6	4
3	4	12	9
6	8	19	14

There is a set of two equations which uses the summed values in the work table to calculate values for a and b. To use these equations, we need a special symbol to express the sums in shorthand form. This is the capital Greek letter sigma (Σ). The symbol is read as "sum of." Thus, Σx is the sum of all the x-values, Σy is the sum of all the y-values, Σxy is the sum of all the xy-values, and Σx^2 is the sum of all the x^2-values.

We also need a symbol to represent the number of known data points and we will denote this with the capital letter N. In the example, N equals 3 (since there are 3 points).

The equations that we are going to use to find a and b are:

$$\Sigma y = Na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

(These equations are known as the "normal equations" and their derivation may be found in any basic statistics text.)

To use these equations, values from the work table are substituted at the appropriate places. From our sample table we have:

$$\begin{aligned} \Sigma x &= 6 \\ \Sigma y &= 8 \\ \Sigma xy &= 19 \\ \Sigma x^2 &= 14 \\ N &= 3 \end{aligned}$$

Substituting these values into the normal equations, we get:

$$8 = 3a + 6b \qquad (1)$$

and

$$19 = 6a + 14b \qquad (2)$$

Now we have two equations with two unknowns (a and b) and to solve them, we use the technique of simultaneous solution. For those whose memory of this technique is dim, what follows is a short refresher course.

The first step in simultaneous solution of two equations is to eliminate one of the unknown terms. This can be done if the coefficients of (i.e., numbers preceding) one of the unknowns are made equal in both equations. This will allow these terms to cancel out when one equation is subtracted from the other. In our example, this can be done easily by multiplying both sides of the equation (1) by 2. When we do this, we get:

$$16 = 6a + 12b \qquad (3)$$

Writing equations (3) and (2) together, we have:

$$\begin{aligned} 16 &= 6a + 12b \qquad (3) \\ 19 &= 6a + 14b \qquad (2) \end{aligned}$$

If equation (2) is subtracted from equation (3), the a-terms cancel out and we can solve for b:

$$\begin{array}{rcr} 16 &=& \cancel{6a} + 12b \\ \ominus 19 &=& \ominus\cancel{6a} + \ominus 14b \\ \hline -3 &=& -2b \\ -3/-2 &=& b \\ 1.5 &=& b \end{array}$$

To get the value of a, we can substitute the computed value of b into either equation (1) or (2). Let's use the first one:

$$\begin{aligned}
8 &= 3a + 6b \qquad\qquad (1)\\
8 &= 3a + 6(1.5)\\
8 &= 3a + 9\\
8 - 9 &= 3a\\
-1 &= 3a\\
-1/3 &= a\\
-0.33 &= a
\end{aligned}$$

So our straight line equation is:

$$y = -0.33 + 1.5x$$

What kind of line does this give us? We can find out by substituting each value of x into the equation to see what values of y we obtain.

For x = 1,

$$\begin{aligned}
y &= -0.33 + 1.5(1)\\
&= -0.33 + 1.5\\
&= 1.17
\end{aligned}$$

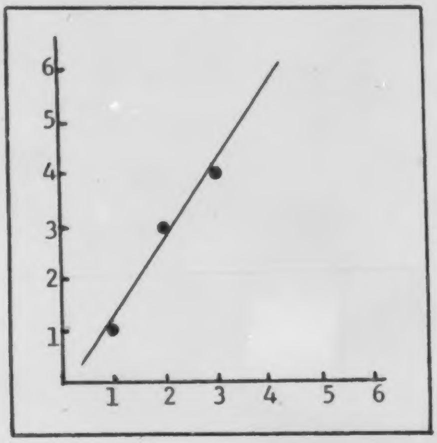
For x = 2,

$$\begin{aligned}
y &= -0.33 + 1.5(2)\\
&= -0.33 + 3\\
&= 2.67
\end{aligned}$$

For x = 3,

$$\begin{aligned}
y &= -0.33 + 1.5(3)\\
&= -0.33 + 4.5\\
&= 4.17
\end{aligned}$$

Now we can plot these y-values on the same graph as our three points and see how the line fits the points:



The process which we have just gone through is known as the linear least squares technique. And the line we found is the least squares line. We will learn more about the properties of this line later.

There is an alternative to using the normal equations to find a and b. This alternative is known as the linear fit algorithm and consists of two equations which can be directly solved for a and b. They are:

$$\boxed{\begin{array}{c} b = \dfrac{\Sigma x \Sigma y - N\Sigma xy}{(\Sigma x)^2 - N\Sigma x^2} \\[2ex] \text{and} \\[2ex] a = \dfrac{\Sigma y - b\Sigma x}{N} \end{array}}$$

These formulas are simply the result of directly solving the normal equations simultaneously for a and b.

Using these equations and our sample work table, we would get:

$$b = \frac{(6)(8) - 3(19)}{(6)^2 - 3(14)}$$

$$= \frac{48 - 57}{36 - 42}$$

$$= \frac{-9}{-6}$$

$$b = 1.5$$

$$a = \frac{8 - 1.5(6)}{3}$$

$$= \frac{8 - 9}{3}$$

$$= -1/3$$

$$a = -0.33$$

And these are the same answers we got before.

Once the least squares equation has been found, it can be used to calculate estimated y-values for other than the known x-values. For example, let's find the y-values associated with x-values of 1 1/2 and 5.

For $x = 1.5$,

$$\begin{aligned} y &= -0.33 + 1.5(1.5) \\ &= -0.33 + 2.25 \\ &= 1.92 \end{aligned}$$

For $x = 5$,

$$\begin{aligned} y &= -0.33 + 1.5(5) \\ &= -0.33 + 7.5 \\ &= 7.17 \end{aligned}$$

This capability of using the least squares equation to get the y-values given by the least squares line is known as *iteration* and will come in handy later.

Properties of the Linear Squares Line

The stated purpose at the beginning of this chapter was to find the best line to fit a given set of data points. This type of approach was needed since, in almost all cases, no one straight line will pass through every given point.

The linear least squares line is that best line. It is the one line which, on the average, comes closest to each of the given points. The mathematical definition of this line is:

> *The linear least squares line is that line which minimizes the sum of the squared differences from the given points to the line.*

For any given set of data points, there is only one line which will fit this definition. Thus,

> *There is one and only one linear least squares line for a given set of data points.*

To illustrate what is meant by "differences," consider the example from the previous section. There, we fit a linear least squares line to the points:

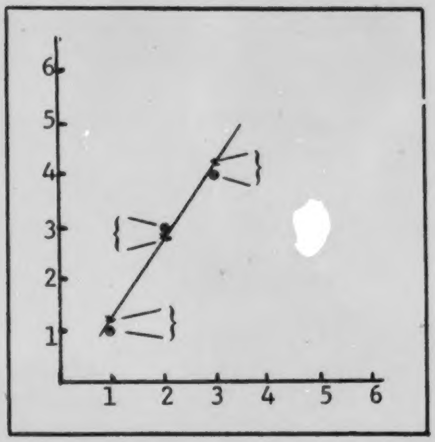
x	y
1	1
2	3
3	4

When we iterated the resulting equation, we got these values for the line (using the symbol y^1 to represent values on the line):

x	y^1
1	1.17
2	2.67
3	4.17

The differences referred to in the definition of the linear least squares line are the differences between the y-values and the y^1-values--i.e., $y-y^1$--at each value of x.

Visually, these differences can be represented by the vertical distances marked by brackets on the following graph:



Mathematically, these differences are:

For $x = 1$,
$y - y^1 = 1 - 1.17 = -0.17$

For $x = 2$,
$y - y^1 = 3 - 2.67 = 0.33$

For $x = 3$,
$y - y^1 = 4 - 4.17 = -0.17$

These differences are also called deviations and are symbolized by the small letter "d."

The "sum of the squared" differences (or deviations) referred to in the definition of the least squares line are calculated by:

- Squaring each of the above deviations; and
- Adding them together.

If we would do this using our example, we would get:

$$(-0.17)^2 + (0.33)^2 + (-0.17)^2 = 0.1667$$

The definition for the linear least squares line refers to this figure (0.1667). The linear least squares line is that line for which this number is the smallest. In other words, any other line fitted to the same points would give a larger number for the sum of the squared deviations.

We can symbolize the sum of the squared deviations as Σd^2 (remember that Σ = sum of).

These deviations can be used to calculate the degree to which the given data points are scattered around the linear least squares line. If we divide the sum of the squared deviations (Σd^2) by the number of known data points less one (N-1), we get the variance. This statistic is a measure of the discrepancy between the actual y-values and the y^1-values.

Another such measure can be found by taking the square root of the variance. The result of this calculation is known as the standard error of the estimate. (This measure may also be referred to as a standard deviation.)

The mathematical formulas for these two statistics can be summarized as follows:

Let s = standard error
s^2 = variance
Σd^2 = the sum of the squared deviations from the linear least squares line
N = the number of known data points

Then,

$$s^2 = \frac{\Sigma d^2}{N-1}$$

$$s = \sqrt{\frac{\Sigma d^2}{N-1}}$$

To aid in the calculation of these measures, it is helpful to set up a work table using the following headings:

$\underline{x}$ $\underline{y}$ $\underline{y^1}$ $\underline{d}$ $\underline{d^2}$

(The first two columns will contain the x- and y-values of the given data points. The third column will list the line-values for each x. Column 4 contains the differences between y and y^1 and column 5 lists these differences squared.)

Using our example, such a work table would look like this:

	x	y	y^1	d	d^2
	1	1	1.17	-0.17	0.0289
	2	3	2.67	0.33	0.1089
	3	4	4.17	-0.17	0.0289
N = 3				Σd^2 =	0.1667

Then,

$$s^2 = \frac{\Sigma d^2}{N-1} = \frac{0.1667}{2} = 0.0834$$

$$s = \sqrt{\frac{\Sigma d^2}{N-1}} = \sqrt{0.0834} = 0.2887$$

Both of these statistics are in the same units as the y-values.

Other properties of the linear least squares line can be stated using the variance and the standard error:

1. The linear least squares line is that line for which the variance between the given points and the line is at its smallest. Conversely, the variance from the given set of points to any other line will be larger.

2. The less scatter there is between a given set of points and its least squares line, the smaller is the value of the standard error. This property can be simply stated by saying that the closer the points are to the line, the less error there will be.

3. When the points being fitted represent a truly linear relationship between the x- and y-values, the deviations about the linear least squares line form a normal distribution.

This third point requires some elaboration. First, it should be noted that not every set of points is linearly-related--i.e., not every set of points is best fitted by a straight line. Some will be better fitted directly by a curved line. Others will require some type of data transformation to bring them closer to a linear relationship. (One example of such data transformation will be given in the next chapter.) The choice of a fitting method should be the result of careful study of the known data.

However, when a set of points _is_ linearly-related, approximately 68% of the deviations from the linear least squares line will have a value which is between ±s (s = one standard error). Approximately 95% of them will be between ±2s and approximately 99% will be between ±3s. This type of an arrangement of data is known as a normal distribution and we will learn more about it in later chapters.

Now it is time to take what has been learned about a least squares line and actually analyze retention data.

CHAPTER 4

THE NATURE OF THE RETENTION CURVE

Data Transformations

We now want to apply the least-squares straight-line technique to longitudinal retention data. To do this, we need to bridge the gap between the retention curve in Figure 4-2 and the retention line on the nomograph (Figure 4-1) on page 58.

In general, when attempting to apply least squares techniques to a curved line, you first determine the mathematical equation that best represents the curve in question. Then it is necessary to determine what can be done mathematically to this equation to get data which are linear.

This type of sequence was followed with the longitudinal retention curve. We tested various known types of curves against the retention curve. From these tests, we learned that the retention curve is best represented by what is known as a log-probability curve equation.

This discovery in itself left us with a rather complex mathematical equation. Fortunately, there are ways to convert this type of equation into a straight line form. These "ways" consist of what are known as data transformations.

A data transformation involves performing the same mathematical operation on each data item (such as taking the square root, squaring, taking reciprocals, etc.). A transformation may be performed on both the x- and y-values or on just the x- or just the y-values. A linear least squares line is then fitted to the transformed data points, the equation for the line is iterated and the resulting line-values are changed back to their original form.

Another look at the line on the nomograph suggests that there are transformations which can be made on retention data to bring out a linear relationship. In this case, both the x- and y-values are involved.

As was indicated in the nomograph chapter, the x-and y-values for retention data are:

x : time since hire
y : % remaining from an original group of hires at time x

The name "log-probability" suggests the data transformations which can be made on these values.

Table 4-1

TABLE OF LOGARITHMS

YEAR	LOG (10)	YEAR	LOG (10)	YEAR	LOG (10)
0.1	-1.00000	5.1	0.70757	10.1	1.00432
0.2	-0.69897	5.2	0.71600	10.2	1.00860
0.3	-0.52288	5.3	0.72428	10.3	1.01284
0.4	-0.39794	5.4	0.73239	10.4	1.01703
0.5	-0.30103	5.5	0.74036	10.5	1.02119
0.6	-0.22185	5.6	0.74819	10.6	1.02531
0.7	-0.15490	5.7	0.75587	10.7	1.02938
0.8	-0.09691	5.8	0.76343	10.8	1.03342
0.9	-0.04576	5.9	0.77085	10.9	1.03743
1.0	0.00000	6.0	0.77815	11.0	1.04139
1.1	0.04139	6.1	0.78533	11.1	1.04532
1.2	0.07918	6.2	0.79239	11.2	1.04922
1.3	0.11394	6.3	0.79934	11.3	1.05308
1.4	0.14613	6.4	0.80618	11.4	1.05690
1.5	0.17609	6.5	0.81291	11.5	1.06070
1.6	0.20412	6.6	0.81954	11.6	1.06446
1.7	0.23045	6.7	0.82607	11.7	1.06819
1.8	0.25527	6.8	0.83251	11.8	1.07188
1.9	0.27875	6.9	0.83885	11.9	1.07555
2.0	0.30103	7.0	0.84510	12.0	1.07918
2.1	0.32222	7.1	0.85126	12.1	1.08279
2.2	0.34242	7.2	0.85733	12.2	1.08636
2.3	0.36173	7.3	0.86332	12.3	1.08991
2.4	0.38021	7.4	0.86923	12.4	1.09342
2.5	0.39794	7.5	0.87506	12.5	1.09691
2.6	0.41497	7.6	0.88081	12.6	1.10037
2.7	0.43136	7.7	0.88649	12.7	1.10380
2.8	0.44716	7.8	0.89209	12.8	1.10721
2.9	0.46240	7.9	0.89763	12.9	1.11059
3.0	0.47712	8.0	0.90309	13.0	1.11394
3.1	0.49136	8.1	0.90849	13.1	1.11727
3.2	0.50515	8.2	0.91381	13.2	1.12057
3.3	0.51851	8.3	0.91908	13.3	1.12385
3.4	0.53148	8.4	0.92428	13.4	1.12710
3.5	0.54407	8.5	0.92942	13.5	1.13033
3.6	0.55630	8.6	0.93450	13.6	1.13354
3.7	0.56820	8.7	0.93952	13.7	1.13672
3.8	0.57978	8.8	0.94448	13.8	1.13988
3.9	0.59106	8.9	0.94939	13.9	1.14301
4.0	0.60206	9.0	0.95424	14.0	1.14613
4.1	0.61278	9.1	0.95904	14.1	1.14922
4.2	0.62325	9.2	0.96379	14.2	1.15229
4.3	0.63347	9.3	0.96848	14.3	1.15534
4.4	0.64345	9.4	0.97313	14.4	1.15836
4.5	0.65321	9.5	0.97772	14.5	1.16137
4.6	0.66276	9.6	0.98227	14.6	1.16435
4.7	0.67210	9.7	0.98677	14.7	1.16732
4.8	0.68124	9.8	0.99123	14.8	1.17026
4.9	0.69020	9.9	0.99564	14.9	1.17319
5.0	0.69897	10.0	1.00000	15.0	1.17609

Time Since Hire: The horizontal distances on the nomograph indicate that the transformation made on the x- (or time) values altered the spacing between successive years. Notice how the large vertical lines get closer and closer together as the number of years gets higher. This type of spacing is called a logarithmic progression.

Without going into the gory mathematical details, we can just say that a logarithmic progression is formed by taking the logarithms of the time values. Every number greater than zero has its own logarithm. The value of the logarithm for any number can be found in tables in most math books, by pressing a key on a calculator, or by using a standard computer function.

For use in retention analysis, you will probably need access to only a small range of logarithmic values. Most of these can be found in Table 4-1. Intermediate values can be found by interpolation.

The standard notation for the logarithm (t) of a number r is:

$$t = \log r$$

The importance of logarithms in retention analysis is due to the effect they have on the distances between numbers. To show this, consider first the logs of 1 and 2.

$$\log 1 = 0$$
$$\log 2 = 0.301$$

Thus, the distance between 1 and 2 (in logarithms) is 0.301. In addition,

$$\log 4 = 0.602$$

Thus, the logarithmic distance from 2 to 4 is:

$$\log 4 - \log 2 = 0.602 - 0.301 = 0.301$$

But this is the same as the logarithmic distance between 1 and 2. This is why on the nomograph, the horizontal distance between "1st year" and "2nd year" equals the horizontal distance between "2nd year" and "4th year."

Similarly, the logarithmic distance from 10 to 100 is the same as that from 1 to 10. What this means is that the logarithms of successively larger numbers are closer and closer together.

It is also possible to transform the logarithm of a number back to the number itself. This can be done by using a table of logs backwards.

Mathematically, if

$$t = \log r$$

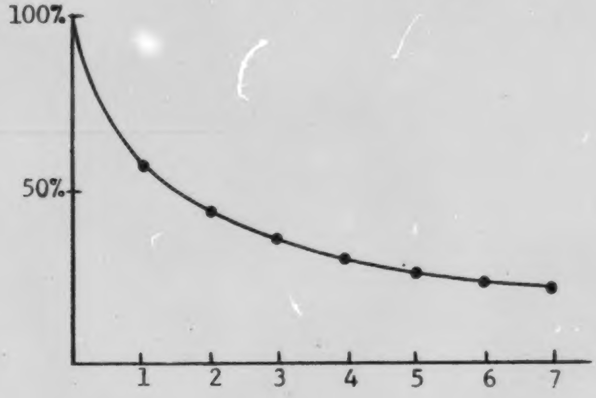
then,

$$r = 10^t$$

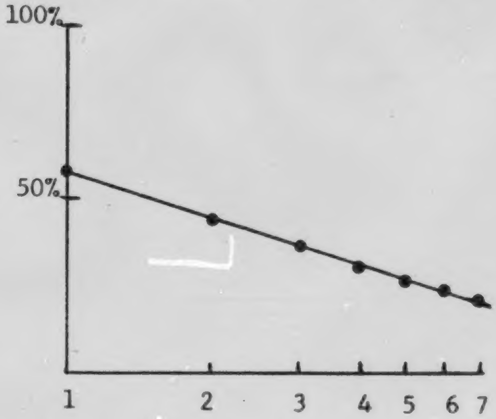
(i.e., 10 raised to the power of t. This is true because we are using what are known as "base 10" logarithms.)

What happens when turnover data are plotted on a standard graph using a logarithmic scale on the x-axis?

Consider again the sample nomograph data. You will remember that, when plotted on a regular graph, the turnover curve looks like this:



Plotting this same data with a logarithmic scale on the x-axis, i.e., with the time values in logs[1] (and no change on the y-axis), we have:

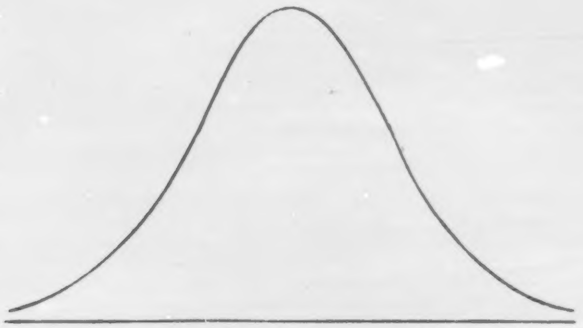


---

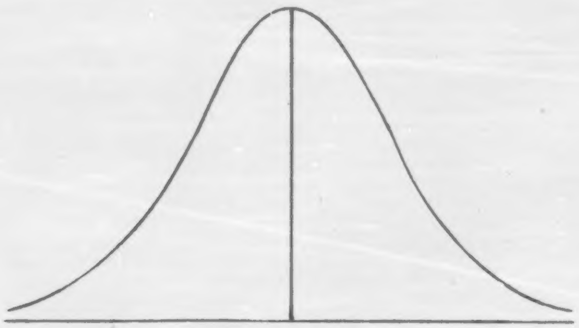
[1] The log scale on the x-axis starts with one (1) since the logarithm of [illegible] is zero.

When retention data are plotted in this manner, their relationship obviously comes much closer to linearity. However, we can get an even better linear relationship by also making a transformation of the y-values (percent retained).

Percent Retained: The second transformation uses certain properties of a normal probability curve. A normal curve (also known as a bell curve) looks like this:



This type of a curve has a number of special properties. First, it is symmetrical about a midline drawn from its highest point perpendicularly to the x-axis, like so:



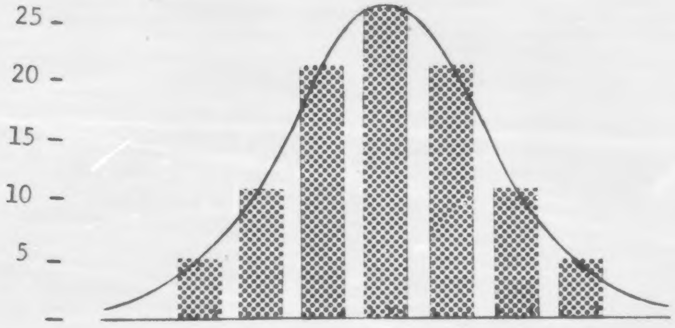
This midline is known as the mean and it splits the area under the normal probability curve into two exact halves. Thus, half the curve is to the left of the mean and half is to the right.

Second, distances from the mean are measured in standard deviations. A standard deviation is the same kind of measure as the standard error which was discussed in the previous chapter. The standard error is a measure of the scatter about a linear least squares line. The standard deviation is a measure of the scatter about the arithmetic mean of a group of values.

If a given group of values is what is known as normally distributed, the plot of the percent of the frequency of occurrence (or probability) of each individual deviation from the mean will form a normal curve. For example, suppose we have the following group of values: 7, 8, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 11, 11, 11, 11, 12, 12, 13. The mean (or average) of these values is 10. The deviations from the mean are found by subtracting 10 from each value. Since most of the values occur more than once, we get the following set of deviations and frequencies (of each deviation):

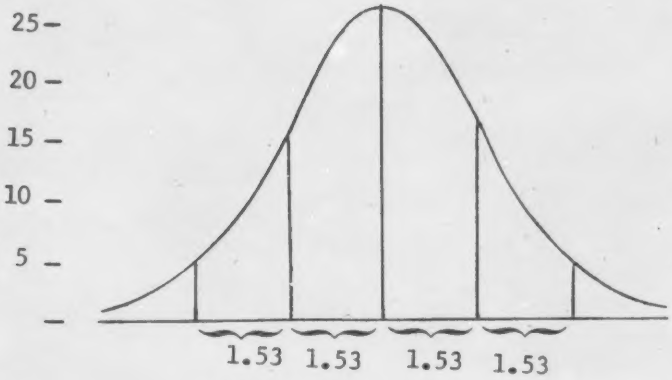
Deviation	Number of Occurrences	Percent of Occurrence
-3	1	5.26%
-2	2	10.53
-1	4	21.05
0	5	26.32
1	4	21.05
2	2	10.53
3	1	5.26
	19	100.00%

Plotting the percent of occurrences using a bar graph, we have:

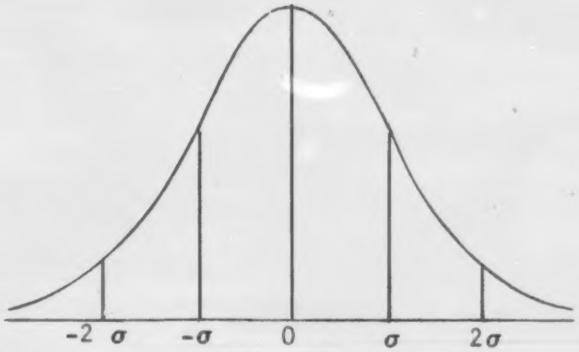


You can see that when a smooth curve is drawn connecting the midpoints of the top of each bar, a normal curve appears.

The standard deviation of these deviation values, using the same formula that was used for standard error ($\sqrt{\Sigma d^2/N-1}$) is 1.53. It was stated previously that distances from the mean of the normal curve are measured in standard deviations. If we re-plot the normal curve above marking off standard deviation distances, we have:



The mathematical symbol for the standard deviation is the small Greek letter sigma (σ). A more general drawing of a normal curve would be:



Note that the values to the left of the mean are negative and those to the right are positive. When talking about or measuring x-axis values for a normal curve, we are concerned with the number of standard deviations from the mean. For example, we might want to know the curve value for a point 1½ σ's from the mean, or $-.37\sigma$, or 3σ's, etc.

As with the standard error, percentage values can be associated with standard deviation values. In this case, the percentage values refer to the percent of the area under the normal curve which is between perpendicular lines drawn from two standard deviation values on the x-axis to the curve. The percent of area under the entire curve, from the leftmost to the rightmost point, is 100%.

The percentage of area between (a) $-\sigma$ and $+\sigma$ is 68%, (b) -2σ and $+2\sigma$ is 95.5% and (c) -3σ and $+3\sigma$ is 99%. These percentages are the same as those between the same ranges of the standard error. (Thus, the deviations from the linear least squares line are normally distributed.) These percentages can also be expressed as probabilities. For example, the probability that a deviation will fall between $\pm\sigma$ is 0.68. In other words, there are 68 chances out of 100 that a deviation will fall

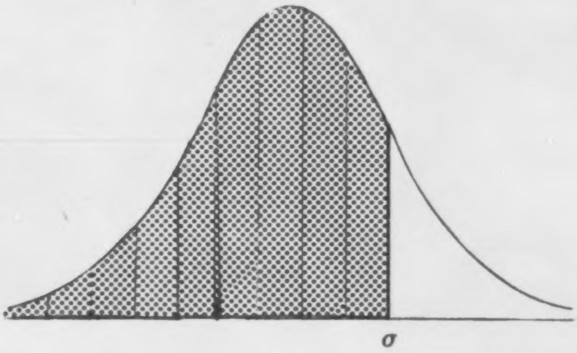
Table 4-2

NORMAL CURVE AREA CONVERSION TABLE

PERCENT REMAINING	STANDARD DEVIATIONS FROM MEAN	PERCENT REMAINING	STANDARD DEVIATIONS FROM MEAN	PERCENT REMAINING	STANDARD DEVIATIONS FROM MEAN
95	1.64485	65	0.38532	35	-0.38532
94	1.55477	64	0.35846	34	-0.41246
93	1.47579	63	0.33185	33	-0.43991
92	1.40507	62	0.30548	32	-0.46770
91	1.34076	61	0.27932	31	-0.49585
90	1.28155	60	0.25335	30	-0.52440
89	1.22653	59	0.22755	29	-0.55338
88	1.17499	58	0.20189	28	-0.58284
87	1.12639	57	0.17637	27	-0.61281
86	1.08032	56	0.15097	26	-0.64335
85	1.03643	55	0.12566	25	-0.67449
84	0.99446	54	0.10043	24	-0.70630
83	0.95417	53	0.07527	23	-0.73885
82	0.91537	52	0.05015	22	-0.77219
81	0.87790	51	0.02507	21	-0.80642
80	0.84162	50	0.00000	20	-0.84162
79	0.80642	49	-0.02507	19	-0.87790
78	0.77219	48	-0.05015	18	-0.91537
77	0.73885	47	-0.07527	17	-0.95417
76	0.70630	46	-0.10043	16	-0.99446
75	0.67449	45	-0.12566	15	-1.03643
74	0.64335	44	-0.15097	14	-1.08032
73	0.61281	43	-0.17637	13	-1.12639
72	0.58284	42	-0.20189	12	-1.17499
71	0.55338	41	-0.22754	11	-1.22653
70	0.52440	40	-0.25335	10	-1.28155
69	0.49585	39	-0.27932	9	-1.34076
68	0.46770	38	-0.30548	8	-1.40507
67	0.43991	37	-0.33185	7	-1.47579
66	0.41246	36	-0.35846	6	-1.55477
				5	-1.64485

between $\pm\sigma$.

In addition, we can associate with every distance (or standard deviation) value from the mean (plus and minus) a unique percentage value which represents the percent of area under the normal curve from the leftmost point to a line drawn perpendicular to the curve from the given standard deviation value. For example, the percent value associated with the mean (or 0σ) is 50% (since exactly half of the curve lies to the left of the mean). Similarly, the percent value associated with $+\sigma$ is 84%. The percent value being measured is shown in the shaded area:



In the same way, one specific percent value can be associated with each standard deviation value. And, conversely, one standard deviation value can be associated with each percent value. (Thus, this is also a two-way transformation.)

This is our second data transformation: transforming "percent remaining"--i.e., percent of employees still remaining from the starting group--to "number of standard deviations from the mean of a normal curve."

As with logarithms, these values can be found in mathematical tables. A simple one is found in Table 4-2. This table requires interpolation for intermediate values. More detailed tables can be found in statistics books. There is also a mathematical formula which may be used. (See the Technical Analysis for LOGPRO in the Appendix section.)

Using the Normal Curve Area Conversion Table (Table 4-2), we can see that a percent remaining figure of 85% corresponds to 1.03643 standard deviations, 80% to 0.84162σ , etc.

Look again at the vertical lines on the nomograph (p. 16) and consider their scaling. The percents on the scale are spaced according to their corresponding standard deviation distance from the mean (which is represented by the 50% mark).

For example, +1 σ corresponds to 84.1% and ½ σ to 69.1%. If a mark is made at each of these values, it can be seen that the distance from the 84.1% point to the 69.1% point is exactly the same (i.e., ½ σ) as the distance from the 69.1% point to the mean (50%). These are not equal distances in straight percent since:

$$\begin{aligned} 84.1\% - 69.1\% &= 15\% \\ 69.1\% - 50.0\% &= 19.1\% \end{aligned}$$

The transformation of "percent remaining" to number of standard deviations from the mean of a normal curve, when used with the transformation of time to log of time, gives the straight line which can be found on the nomograph.

Now we have the two data tranformations that were needed:

1. Years of service is transformed into the logarithm of years of service; and

2. Percent of employees retained from an original group of hires is transformed into number of standard deviations from the mean of a normal probability curve.

Some Comments About Retention Analysis

These comments will cover four areas:

1. The collection of retention data;
2. The reasons behind the close relationship between retention data and the log-probability technique;
3. The condition under which the projection of a log-probability equation remains valid; and
4. The continuity of the retention curve.

Data Collection

It has been stated that retention data are collected by following a group of hires over time and counting the number left after certain lengths of time.

The group that is followed (the "original group" or "cohort") is composed of employees hired during a specific span of time. This span should be limited to one year or less. The "certain lengths of time" could be any standard time units (years, months, days) after the end of the original time span. Since the original group will contain employees with different lengths of service at the end of the original time span (from 1 day to 1 year), there are several options available for assigning values to the time units.

For example, consider a group of employees hired during a given fiscal year ending September 30. And suppose that retention data for this group are recorded on succeeding September 30th's. The time values for these data points may be:

(a) 1.0 year, 2.0 years, etc. (calculated from September 30th to September 30th); or

(b) 1.5 years, 2.5 years, etc. (assuming an even distribution of hires during a year and using the midpoint; i.e., 0.5); or

(c) 1.x years, 2.x years, etc. (where x is a factor based upon specific hiring patterns, such as more hiring done at the end of a fiscal year); or

(d) 1.y years, 2.y years, etc. (where y represents the actual average length of service of the original group at the end of the hiring time span).

The number to the right of the decimal point is known as the time averaging factor. When this factor is greater than 0, it represents the distribution of hiring into a group during a time span.

Different occupations have different turnover/retention patterns. Some occupations, such as Internal Revenue Agent, Foreign Service Officer, etc., are unique to government. This type of occupation does not exist in the private sector. Consequently, since there are no places other than government where the employee can go and still remain in his or her occupation, the turnover in these occupations tends to be low. Other occupations, such as Clerk/Typist, are widely distributed both inside and outside of government. Thus, there are many possible employers for people in these jobs. Consequently, the turnover in these occupations tends to be high.

These differences in turnover patterns among occupations require that retention data be collected separately by occupation. It may be possible to combine occupations with similar turnover patterns at a later stage of retention analysis. However, the original data should still be collected separately.

It may also be desirable to collect retention data by grade-at-hire within an occupation since there may be some variation in the turnover patterns of employees in different entry grade levels (e.g., GS-5 vs GS-7).

Why Retention Is A Log-Probability Function

Our analysis of longitudinal retention data led us to the use of a log-probability equation. Why does this particular equation fit retention data so well?

The concentration of turnover in the early years of service suggests that turnover behavior is exhibited logarithmically over time. As with the values of logarithms, the figures for percent retained from an original group get closer together as the number of years since hire gets higher.

A logarithmic progression over time is also characterized by equal movements during equal logarithmic times. For example, consider the data for the printed line on the nomograph (Figure 2-2 , p. 16).

Losses from 1st year to 2nd year = 58% - 44% = 14%
Losses from 2nd year to 3rd year = 44% - 36% = 8%
Losses from 3rd year to 4th year = 36% - 30% = 6%

Note that the percent lost from the 1st year to the 2nd year (14%) is equal to the percent lost from the 2nd year to the 4th year (8% + 6% = 14%). (And remember that the logarithmic distance from 1 to 2 equals that from 2 to 4.) This also suggests that turnover behavior is displayed logarithmically over time.

The effectiveness of the second, or "probability," tranformation relates to the distribution of attitudes toward work which is displayed in most groups. This distribution tends to form a normal probability curve.

Each person's attitude toward his or her job is made up of literally hundreds of elements and combinations. For even a small group, then, the total number of such elements will reach into the thousands. For only a few employees are these attitudes likely to be extremely negative. Likewise, for only a few employees are they likely to be extremely positive. By far the likeliest is that an employee's set of work attitudes will tend to concentrate toward the midpoint between these two extremes (negative and positive), where the positive and negative factors are more nearly in balance. This distribution of employee work attitudes tends to form a normal probability curve.

In terms of group losses we might express this by saying that in any given group of hires there is likely to be: one small group that will hate the job and want to quit immediately; one small group that will love the job and never want to leave; but the bulk of the group of hires, who have some positive and some negative feelings about the job, will be bunched up in between. Their turnover will likely be neither early nor late, but spread out over a substantial period of time.

Condition For Projection Validity

For log-probability projections of longitudinal retention data to remain valid it is necessary that the organization's internal situational factors affecting a group's retention behavior during the initial period of empirical observation continue substantially unchanged throughout the period of the projection.

This condition involves the total working situation within a given organization. As long as this working situation does not go through any major changes (such as a large reorganization, a radical program change, etc.) during the data collection and projection periods, turnover patterns will not change. However, major changes in the working situation will be followed by changes in employee turnover. This is because such changes affect the mix of positive and negative work attitudes within the groups touched by the changes. Some employees who felt positive toward the previous work situation will feel negative toward the new one and vice versa. After an initial period of adjustments, turnover will again settle into a log-probability pattern albeit a somewhat different one than before.

Continuity of a Retention Curve

One of the most interesting characteristics of the retention curve is its relative independence of external economic and political happenings.

For example, one study of retention followed a group of hires into the Federal service in the most populous professional, administrative, and technical occupations during FY-63. The same group of employees was followed for 7 years (through 1970) and the percent retained was recorded at the end of each fiscal year. These percents were plotted on a nomograph (Figure 4-1) and the "best" line was estimated using a ruler. Notice how close the fit is.

Figure 4-2 shows the same data plotted arithmetically. A curve was fitted to the points using the log-probability least squares technique. Again, notice how closely the curve fits the points. Note that at no time during the seven-year period do the actual values differ from the fitted curve by more than 1%.

This means that the retention curve for this group continued undisrupted over the entire seven-year period. But this was an era when the private economy was experiencing substantial ups and downs. It was also the era during which the number of Federal employees first greatly increased then began to decrease again due to the situation in Viet Nam. These were truly major changes yet neither of these had any effect on the retention curve of the group.

Figure 4-1

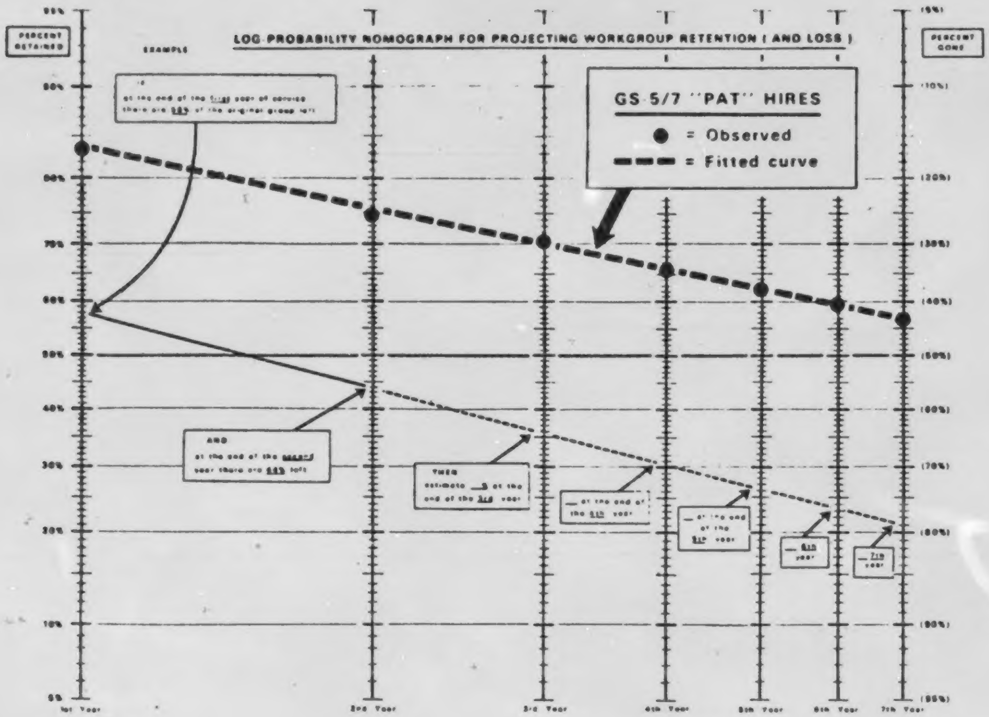
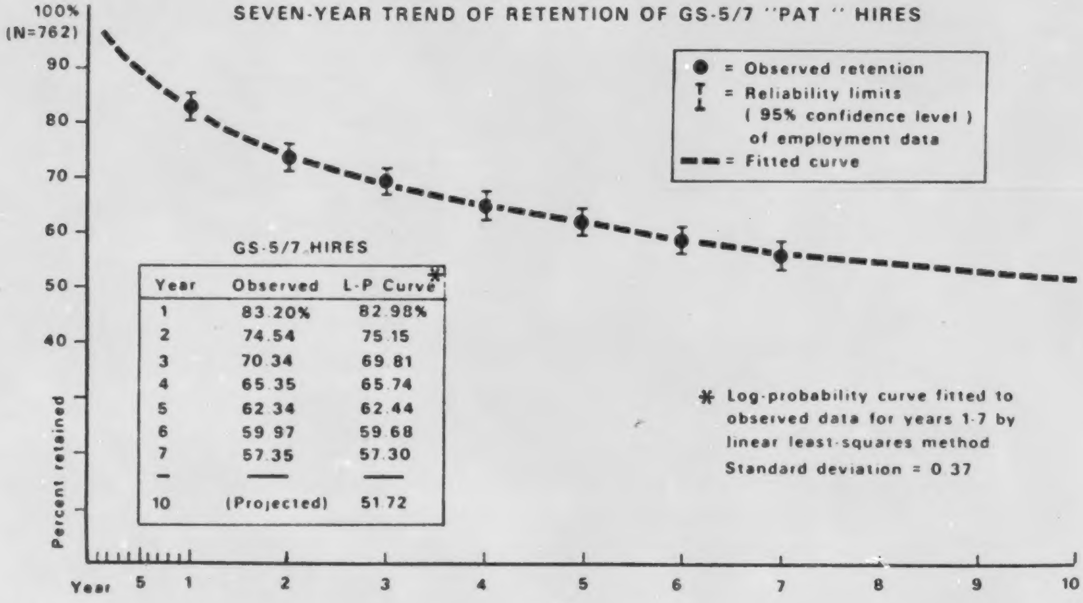


Figure 4-2



GS-5/7 HIRES

Year	Observed	L-P Curve*
1	83.20%	82.98%
2	74.54	75.15
3	70.34	69.81
4	65.35	65.74
5	62.34	62.44
6	59.97	59.68
7	57.35	57.30
—	—	—
10	(Projected)	51.72

* Log-probability curve fitted to observed data for years 1-7 by linear least-squares method
Standard deviation = 0.37

From these and similar findings, we conclude that log-probability type turnover curves, once established, are for practical purposes nonresponsive to external economic and political influences.

CHAPTER 5

FITTING AND USING THE LOG-PROBABILITY CURVE

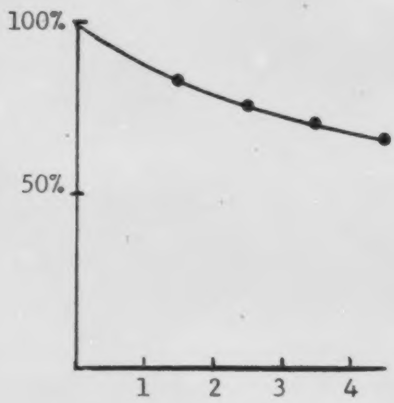
Manual Calculation of the Log-Probability Equation

Now that we know about the necessary data transformation and data collection techniques, we are actually ready to use the method of least squares to fit a log-probability turnover curve. (A more detailed version of this process may be found in the appendix Manual Calculation of the Log-Probability Curve.)

As an example, let's use the following data:

Years Since Hire	Percent Retained[1]
1.5	83.2%
2.5	75.6%
3.5	70.4%
4.5	65.6%

The graph of this retention curve looks like this:



The first step in fitting these data is to convert them into their linear form using the data transformations discussed in Chapter 4. The "Years Since Hire" can be transformed directly into logs using Table 4-1. From this table we get:

---

[1] Note that throughout this and the following chapters an employee is considered "retained" only if the employee continues to be in the same occupational group for which the analysis is made. Any employee who leaves this group in any way (quit, change to other occupation, etc.) is a "loss," even if that employee continues to work in the same organization.

log 1.5 = 0.17609
log 2.5 = 0.39794
log 3.5 = 0.54407
log 4.5 = 0.65321

The second transformation--from percent remaining to number of standard deviations from the mean of the normal curve--can be done by using the Normal Curve Area Conversion Table (Table 4-2, page 52) and some linear interpolation.

Linear interpolation is simply a method to find a value for a number which falls between two entries in a table. For example, what standard deviation value corresponds to 83.2% (our first data point)? Now, 83.2% is two-tenths of the way between 83% and 84%. So we want to find the standard deviation value which is two-tenths of the way between 0.95417 (83%) and 0.99446 (84%). To do this, we can multiply the difference between these two values by 0.2 and then add this difference to the smaller standard deviation value (0.95417). If we do this, we get:

(1) $0.99446 - 0.95417 = 0.04029$
(2) $(0.2) \times (0.04029) = 0.00806$
(3) $0.95417 + 0.00806 = 0.96223$

Thus, the standard deviation value corresponding to 83.2% is 0.96223. Continuing in this manner, we get:

Percent Retained	Standard Deviations
83.2%	0.96223
75.6%	0.69358
70.4%	0.53599
65.6%	0.40160

Now we have a set of x- and y-values and can set up a work table such as the one described in Chapter 3. To help keep the data relationships clear, we will expand the work table to include both the years since hire and percent retained values.

Log-Probability Work Table

Years Since Hire	Log Year (x)	Percent Retained	Stan. Dev. From Mean (y)	(x)(y)	x^2
1.5	0.17609	83.2%	0.96223	0.16944	0.03101
2.5	0.39794	75.6%	0.69358	0.27600	0.15836
3.5	0.54407	70.4%	0.53599	0.29162	0.29601
4.5	0.65321	65.6%	0.40160	0.26233	0.42668
N=4	1.77131		2.59340	0.99939	0.91206

The next step is to substitute values from the work table into either (a) the normal equations or (b) the linear fit algorithm.

As a reminder, the normal equations are:

$$\sum y = Na + b\sum x$$
$$\sum xy = a\sum x + b\sum x^2$$

From the work table, we have:

$$\sum x = 1.77131$$
$$\sum y = 2.59340$$
$$\sum xy = 0.99939$$
$$\sum x^2 = 0.91206$$
$$N = 4$$

Substituting these values into the normal equations gives:

$$2.59340 = 4a + 1.77131b$$
$$0.99939 = 1.77131a + 0.91206b$$

Solving these equations simultaneously gives:

$$a = 1.165275$$
$$b = -1.167328$$

And the log-probability equation is:

$$y = 1.165275 - 1.167328x$$

The linear fit algorithm is:

$$b = \frac{N\sum xy - \sum x\sum y}{N\sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b\sum x}{N}$$

Substitution results in:

$$b = \frac{4(0.99939) - (1.77131)(2.59340)}{4(0.91206) - (1.77131)^2}$$

$$= -1.167328$$

$$a = \frac{2.59340 - (-1.167328)(1.77131)}{4}$$

$$= 1.165275$$

And, again, the log-probability equation is:

$$y = 1.165275 - 1.167328x$$

Now that we have the log-probability equation, what can we do with it?

Iteration and Projection of the Equation

One major use of the log-probability equation is to calculate curve values for given time values. This function includes finding curve values for:

1. The time values used to calculate the equation (so that curve values can be compared with actual values);

2. Any time values which fall between those used to calculate the equation (to estimate retention at intermediate times); and

3. Any future time values (to project numbers to be retained at any future point in time).

For the first two types of time values, this process is called iteration. For the third, it is called projection. But the mathematical steps involved are the same in all three cases. For any given time value, these steps are:

1. Convert the time value to logs.

2. Substitute the log of the time value into the log-probability equation. This substitution involves:

    a. Multiplying the log of the time value by the value of b, and

    b. Adding the result of this multiplication to the value of a.

    The result of this substitution is the curve value given by the equation at the given time value. This curve value is in standard deviations from the mean of a normal curve.

3. Convert the resulting curve value to its corresponding percent remaining value.

For example, using the equation which we calculated in the previous section and the time-value 1.5 years, we get:

$$\begin{aligned} y &= 1.165275 - 1.167328 (\log 1.5) \\ &= 1.165275 - 1.167328 (0.17609) \\ &= 0.95972 \end{aligned}$$

This value is in standard deviation units. Using Table 4-2 and interpolation, it converts to a percent remaining value of 83.14%.

To find the curve value at an intermediate point, say 2.75 years, we go through the same steps:

$$\begin{aligned} y &= 1.165275 - 1.167328 (\log 2.75) \\ &= 1.165275 - 1.167328 (0.43933) \\ &= 0.65243 \end{aligned}$$

This value converts to 74.29%.

To estimate the percent remaining at a future time value, say 5.5 years, we again do the same things:

$$\begin{aligned} y &= 1.165275 - 1.167328 (\log 5.5) \\ &= 1.165275 - 1.167328 (0.74036) \\ &= 0.30103 \text{ (or } 61.83\%\text{)} \end{aligned}$$

To convert these "percent remaining" figures to "number of employees left from the original group," divide the "percent remaining" figure by 100 (to get the percent to its decimal form), and then multiply the number in the original group by this quotient. Suppose, for example, that the original group consisted of 250 employees. Then the estimated number of employees remaining five years since the end of the hiring span (i.e., 5.5 years using the averaging factor) is:

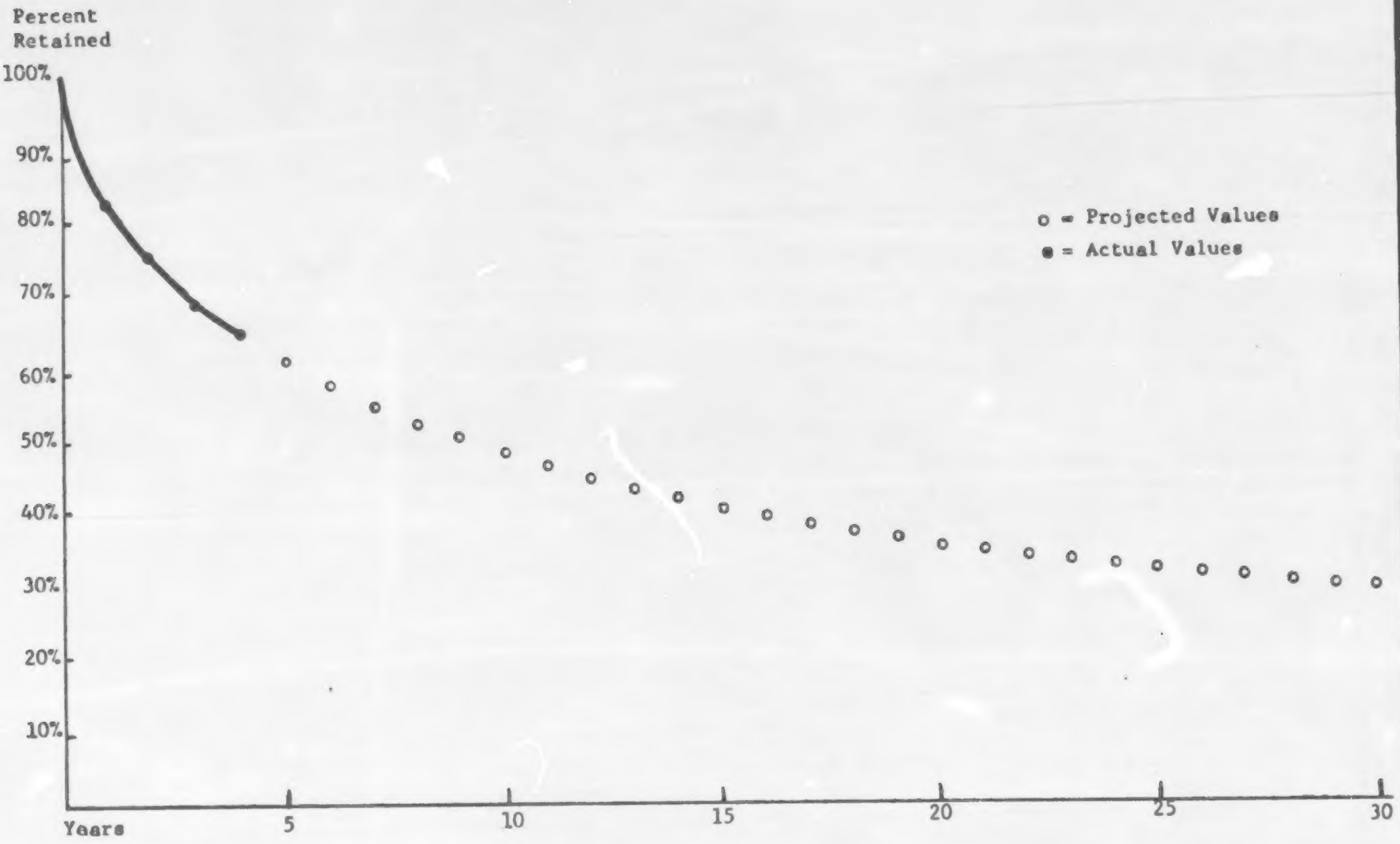
$$\frac{61.83\%}{100} \times 250 = 155 \text{ employees}$$

(Note - If you use a calculator, a mathematical formula, or a computer program to find percent remaining values, the result will be in decimal form so that division by 100 will not be necessary.)

A 30-year projection of the sample log-probability equation is shown in Figure 5-1. This graph also shows the positions of the four actual percent remaining values. Here, again, is the standard retention curve.

The points which are plotted for years 5 to 30 in Figure 5-1 represent the most probable values for projected future retention. They are estimates. It is also possible to determine a range of values within which future retention values will most likely fall. We will show how to do this in an upcoming section.

Figure 5-1


Percent Retained
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
Years
5
10
15
20
25
30
○ = Projected Values
● = Actual Values

Application To Hiring Levels

In Chapter 2, we applied data read from the nomograph to several practical planning situations. We can now use the log-probability equation to handle the same situations. To demonstrate this, let's use the log-probability equation we calculated earlier to represent a given occupation:

$$y = 1.165275 - 1.167328x$$

Suppose, as we did in Chapter 2, that an organization has been hiring exactly 100 employees into a given occupation each year for the last 6 years. Suppose also that this hiring was evenly distributed during a year (so that the averaging factor will be 0.5). To get a length-of-service distribution[2/] of those employees still on board, first iterate the equation using the time values 1.5, 2.5, . . . 6.5 (representing from 1 to 6 years since hire). When we do this, we get:

Year	Percent Remaining
1.5	83.14%
2.5	75.83
3.5	70.20
4.5	65.64
5.5	61.83
6.5	58.56

Translating these percent values to numbers retained (with 100 employees in each original group) gives:

Years Since Hire	Number of Employees Retained
6	59
5	62
4	66
3	70
2	76
1	83
	416

Now we want to determine how many of these 416 employees will leave during the upcoming year. By the end of this next year, each employee still on board will move to the next length-of-service category. For

---

[2/] In an operating situation, a length-of-service distribution can be found by simply counting the number in the given occupation who are in each length-of-service category.

example, those in the one-year-since-hire category will move to the two-years-since-hire category. From the log-probability equation, we know that 75.83% of the original group are retained to the end of the second year. So this group will go from 83 to 76 (100 x 0.7583) employees. Each of the other groups are reduced in the same manner. (For the smallest group, we need the percent retained value for 7.5 years which is 55.72%.)

This process gives the following figures for the end of the next year:

Years Since Hire	Number at End of Current Year	Number One Year Later	Lost in Year
6	59	56	3
5	62	59	3
4	66	62	4
3	70	66	4
2	76	70	6
1	83	76	7
	416	389	27

Thus, 27 of the 416 employees will leave during the upcoming year. This is a turnover rate of .0649 (27/416) or 6.49%.

We can use these figures to show how different levels of hiring affect the turnover rate of a group.

Suppose that during the current year the same pattern of hiring continues; i.e., 100 employees are hired into the occupation in an even distribution during the year. Using the log-probability equation we know that 83 of these 100 hires will be retained by the end of the upcoming year. This means that 17 of them will leave. Combining these losses with those from the other length-of-service categories gives:

17 + 27 = 44 losses

out of

100 + 416 = 516 employees

This gives a turnover rate of .0853 (44/516) or 8.53%.

Now suppose that instead of hiring 100 employees into the occupation this year, the organization decides to hire 200. The number remaining from this group of hires at the end of the upcoming year would be 166 (i. e., 200 x .8314). Thus, 34 of the new hires will be lost. Combining these losses with those from the other length-of-service categories (which remain the same) gives:

$$34 + 27 = 61 \text{ losses}$$

out of

$$200 + 416 = 616 \text{ employees}$$

In this case, the turnover rate would be .0990 (61/616) or 9.90%.

Continuing in the same manner gives the following turnover percentages for the group under different levels of hiring:

Number Hired Current Year	Number Lost	Number of Employees	Turnover Percent
0	27	416	6.49%
50	35	466	7.51
100	44	516	8.53
200	61	616	9.90
300	78	716	10.89

These data illustrate again a point which was discussed in Chapter 2 but which is of such importance that it will be stressed again here. And that is that the rise and fall of turnover rates is in response to the rise and fall of hiring rates.

Thus, when a management decision is made to increase hiring, the impact of the decision on turnover can be determined. By the same token, if a decision is made to freeze hiring or even to have a reduction-in-force, the subsequent decline in turnover can be measured.

Other Applications

The log-probability equation can be applied to other personnel questions. A few of these applications will be discussed here. We will continue to use the same sample log-probability equation.

1. What is the probability that a new hire into an occupation will still be on board x years from now?

This question can be answered simply by iterating the log-probability equation for the value of x. For example, if we want to know the probability that a new hire into our sample group will last 6 years, we iterate the log-probability equation for 6 years:

$$\begin{aligned} y &= 1.165275 - 1.167328 (\log 6) \\ &= 1.165275 - 1.167328 (0.77815) \\ &= 0.25692 \end{aligned}$$

This standard deviation value translates to 60.14% or a probability value of 0.6014. So that there is a 60% chance that a new hire into this occupation will stay for 6 years.

2. What is the probability that an employee who has already been on board for x years will still be on board in t years?

To answer this question, we need two values:

(a) the probability that an employee will stay for x years, and

(b) the probability that an employee will stay for x + t years.

Then the probability that an employee with x years of service will stay for t more years is calculated by dividing (b) by (a).

For example, if we want to know the probability that an employee who has 3 years of service will stay for 3 more years, we first iterate the log-probability equation for the values 3 and 6 (3 + 3). This gives:

$$\begin{aligned} y &= 1.165275 - 1.167328\ (\log 3) \\ &= 1.165275 - 1.167328\ (0.47712) \\ &= 0.60832 \end{aligned}$$

$$\begin{aligned} y &= 1.165275 - 1.167328\ (\log 6) \\ &= 1.165275 - 1.167328\ (0.77815) \\ &= 0.25692 \end{aligned}$$

The first standard deviation value (0.60832) transforms to a probability of 0.7285. The second (0.25692) transforms to 0.6014. So that the probability that an employee will stay 3 years is 0.7285 and the probability that an employee will stay 6 years is 0.6014.

Now the probability that an employee with 3 years of service will stay for 6 years is:

$$\frac{\text{Probability of staying 6 years}}{\text{Probability of staying 3 years}}$$

Or, in this case:

$$\frac{0.6014}{0.7285} = 0.8255$$

Thus, in this group, an employee with 3 years of service has an 83% chance of remaining 3 more years. Or, in other words, if 100 employees in the given occupation have 3 years of service, then 83 of them will stay for at least 3 more years.

3. How many employees must be hired now so that x of them will be on board in t years?

To answer this question, we simply iterate the log-probability equation for t + s years (where s = an averaging factor, if any) and divide the number of employees wanted in t years by the resulting probability.

For example, if we want to have 100 employees in our sample occupation on board in 3 years, how many should be hired this year (supposing that these hires are made evenly throughout the year)? First, we must iterate the log-probability equation for 3.5 years (assuming an averaging factor of 0.5):

$$\begin{aligned} y &= 1.165275 - 1.167328 (\log 3.5) \\ &= 1.165275 - 1.167328 (0.54407) \\ &= 0.53017 \end{aligned}$$

This transforms to a probability value of 0.7020. To get the desired answer, we divide 100 (the number of employees wanted on board in 3 years) by 0.7020. This gives:

$$\frac{100}{.7020} = 142 \text{ employees}$$

Thus, 142 employees must be hired in the occupation this year so that 100 of them will be on board in 3 years.

4. Assuming that a training course costs $1000 per employee, what is the actual cost per employee trained and still on board in t years when (a) only new hires are trained and (b) only employees with x years of service are trained?

Suppose that a decision is made to train 100 employees in the sample occupation. And suppose we want to know the actual cost per employee trained and still on board three years after the training program.

In case (a), all of the employees trained would be new outside hires. Three years after the training program those employees from this group who are still on board will have an average length of service of 3.5 years (assuming an even distribution of hiring during a year).

The first step is to calculate the probability that these employees will be retained in three years. To do this, we use the method for question 1; i.e., iterate the log-probability equation for the time value desired. In this case, the time value is 3.5 years. This gives:

$$\begin{aligned} y &= 1.165275 - 1.167328 (\log 3.5) \\ &= 1.165275 - 1.167328 (0.54407) \\ &= 0.53017 \end{aligned}$$

This translates to a retention probability of 0.7020.

To determine the actual cost of the training program per new hire still on board in 3 years, divide the billed cost of the training ($1000 per employee) by the 3-year retention probability (0.7020). This gives:

$$\frac{\$1000}{0.7020} = \$1425/\text{employee trained}$$

Thus, the 3-year return on the training investment when only new outside hires are trained is $1425 per employee trained and still on board.

For case (b), suppose that all the employees to be trained already have three years of service. Three years after the training program those employees from this group who are still on board will have an average length of service of 6.5 years.

In this case, we need to calculate the probability that employees with 3 years of service will reach 6 years of service. To do this, we use the method of question 2. Using this method gives:

$$\begin{aligned} y &= 1.165275 - 1.167328 (\log 3.5) \\ &= 1.165275 - 1.167328 (0.54407) \\ &= 0.53017 \text{ (or } 0.7020) \end{aligned}$$

$$\begin{aligned} y &= 1.165275 - 1.167328 (\log 6.5) \\ &= 1.165275 - 1.167328 (0.81291) \\ &= 0.21634 \text{ (or } 0.5856) \end{aligned}$$

The retention probability for this group is:

$$\frac{0.5856}{0.7020} = 0.8342$$

Now we can again divide the billed cost of the training program ($1000 per trainee) by the calculated retention probability (0.8342). This gives:

$$\frac{\$1000}{0.8342} = \$1199/\text{employee trained}$$

Thus, the 3-year return on the training investment for this group is $1199 per employee trained and still on board.

The important thing to notice here is the per employee cost difference that occurs solely as the result of a selection decision on who is to be trained. In the long run, it will cost an organization more money to train new hires than to train employees who have been on board for a while. This is a factor which should be taken into account when training decisions are made.

The same type of analysis can be made for any other personnel area involving the selection of employees. This includes decisions on filling vacancies by hiring from the outside or by reassignment from within. The outside hire has a greater probability of leaving in less time than does a reassigned employee. Again, this is something a manager should know before such decisions are made. The manager may still decide to fill vacancies by outside hiring but the higher turnover level that follows should not be a surprise.

More detailed descriptions of the applications discussed in this section plus other applications can be found in the Appendix section.

The applications shown in this section are examples of some of the kinds of analysis which personnel officials should perform to assess a management workforce plan. They can be used to answer such questions as:

- Is it possible to provide the workforce required by the plan?
- Can the workers needed be provided on the time schedule which has been established?
- How much will providing the required workforce cost?

If analysis shows that a workforce plan is in some way infeasible, then management should be (a) informed on why the plan is not feasible and (b) advised on possible adjustments that could be made to it.

Confidence Range for Projected Values

In a previous section, we showed how to use the log-probability equation to obtain estimates of percent retained at future times. Such projections give the most probable values of future retention. In reality, the actual retention values will most likely fall within a range of values surrounding each projected point. This range can be found by using an important property of the standard error; i.e., that 95% of the deviations from a least squares line will fall between -2s and +2s.

This property makes it possible to calculate a range of values into which an actual retention figure will fall 95% of the time. This range is known as the 95%- or confidence range.

Once a log-probability equation has been calculated, the major steps in calculating the 95%-range are:

1. Iterate the log-probability equation for the known points and find the deviations between the actual values and those given by the log-probability line.
2. Compute the standard error.
3. Project the log-probability equation for the values desired.

78

4. For each projected value, calculate the upper and lower limits of the range in standard deviation units (using ± 2s as boundaries).

5. Convert the projected values and the range limits to their corresponding percent remaining values.

6. Convert these percent values to numbers of employees.

To illustrate this process, we fit a least squares line to the following retention data:

Years Since Hire	Percent Remaining
1	58%
2	44%
3	36%
4	30%

Using an averaging factor of 0.5 years and converting the time values to logs and the percent values to number of standard deviations from the mean of the normal curve gives:

Log Year	Stan. Devs. From Mean
0.17609	0.20189
0.39794	-0.15097
0.54407	-0.35846
0.65321	-0.52440

The log-probability equation that results from fitting these data is:

$$y = 0.4638 - 1.5171x$$

Now we are ready to begin the 6-step process of calculating the confidence ranges associated with the projection of this log-probability equation.

Step 1. The deviations from the least squares line represented by the log-probability equation are calculated by:

a. Iterating the equation for each given x-value to get the value given by the line (y^1).

b. Calculating $y-y^1$ for each given x-value.

These calculations are performed with the y and y^1 values in units of standard deviations from the mean of the normal curve; i.e., without transforming them back to percent remaining values.

Going through this process for the first x-value (1.5 years) gives:

$$\begin{aligned} y^1 &= 0.4638 - 1.5171 (\log 1.5) \\ &= 0.4638 - 1.5171 (0.17609) \\ &= 0.19667 \end{aligned}$$

The actual y-value at this point was 0.20189. Thus, the difference between the line and the actual data at x = 1.5 is:

$$y - y^1 = 0.20189 - 0.19667 = 0.00522$$

Each of the other deviations is calculated in the same manner. The process is summarized in the following table (which also includes the square of the differences):

x	y	y^1	$y-y^1 = d$	d^2
0.17609	0.20189	0.19667	0.00522	0.000027
0.39794	-0.15097	-0.13992	-0.01105	0.000122
0.54407	-0.35846	-0.36161	0.00315	0.000010
0.65321	-0.52440	-0.52719	0.00279	0.000008
				0.000167

Step 2. The standard error about this line is:

$$s = \sqrt{\frac{\Sigma d^2}{N-1}} = \sqrt{\frac{0.000167}{3}}$$
$$= \sqrt{0.000056} = 0.00748$$

Step 3. Suppose that we wish to project the equation through 10 years since hire (using x-values of 5.5 to 10.5). This gives:

Years Since Hire	Projected Value (y^1)
5	-0.65940
6	-0.76947
7	-0.86375
8	-0.94622
9	-1.01950
10	-1.08545

Step 4. To calculate the upper limit (U) and the lower limit (L) of a confidence range, we use the formulas:

$$U = y^1 + 2s$$
$$L = y^1 - 2s$$

where y^1 represents a projected value.

In this case, $2s = 2 \times 0.00748 = 0.01496$.

For the first projected value ($y^1 = -0.65940$), the upper and lower limits would be:

$$\begin{aligned} U &= y^1 + 2s \\ &= -0.65940 + 0.01496 \\ &= -0.64444 \end{aligned}$$

And,

$$\begin{aligned} L &= y^1 - 2s \\ &= -0.65940 - 0.01496 \\ &= -0.67436 \end{aligned}$$

Doing this for each projected value gives:

	--------Range (in Stan. Devs.)----------		
Years Since Hire	Upper Limit	Projected Value	Lower Limit
5	-0.64444	-0.65940	-0.67436
6	-0.75451	-0.76947	-0.78443
7	-0.84879	-0.86375	-0.87871
8	-0.93126	-0.94622	-0.96118
9	-1.00454	-1.01950	-1.03446
10	-1.07049	-1.08545	-1.10041

Step 5. So far, we have been working with standard deviation values. Now we can convert each of the standard deviation values to percent (using Table 4-2 and interpolation). This gives:

Years Since Hire	Range (in Percent) Upper Limit	Range (in Percent) Projected Value	Range (in Percent) Lower Limit
5	25.96%	25.48%	25.00%
6	22.53%	22.08%	21.64%
7	19.80%	19.39%	18.98%
8	17.59%	17.20%	16.82%
9	15.76%	15.40%	15.05%
10	14.22%	13.89%	13.56%

Thus, for example, there is a 95% chance that the percent of the original group retained at the end of the fifth year since hire will be between 25.00 and 25.96. In other words, 95 times out of 100 the actual retention percent will fall between these two values.

A graph of this sample retention curve showing the confidence ranges for the projected values is found in Figure 5-2.

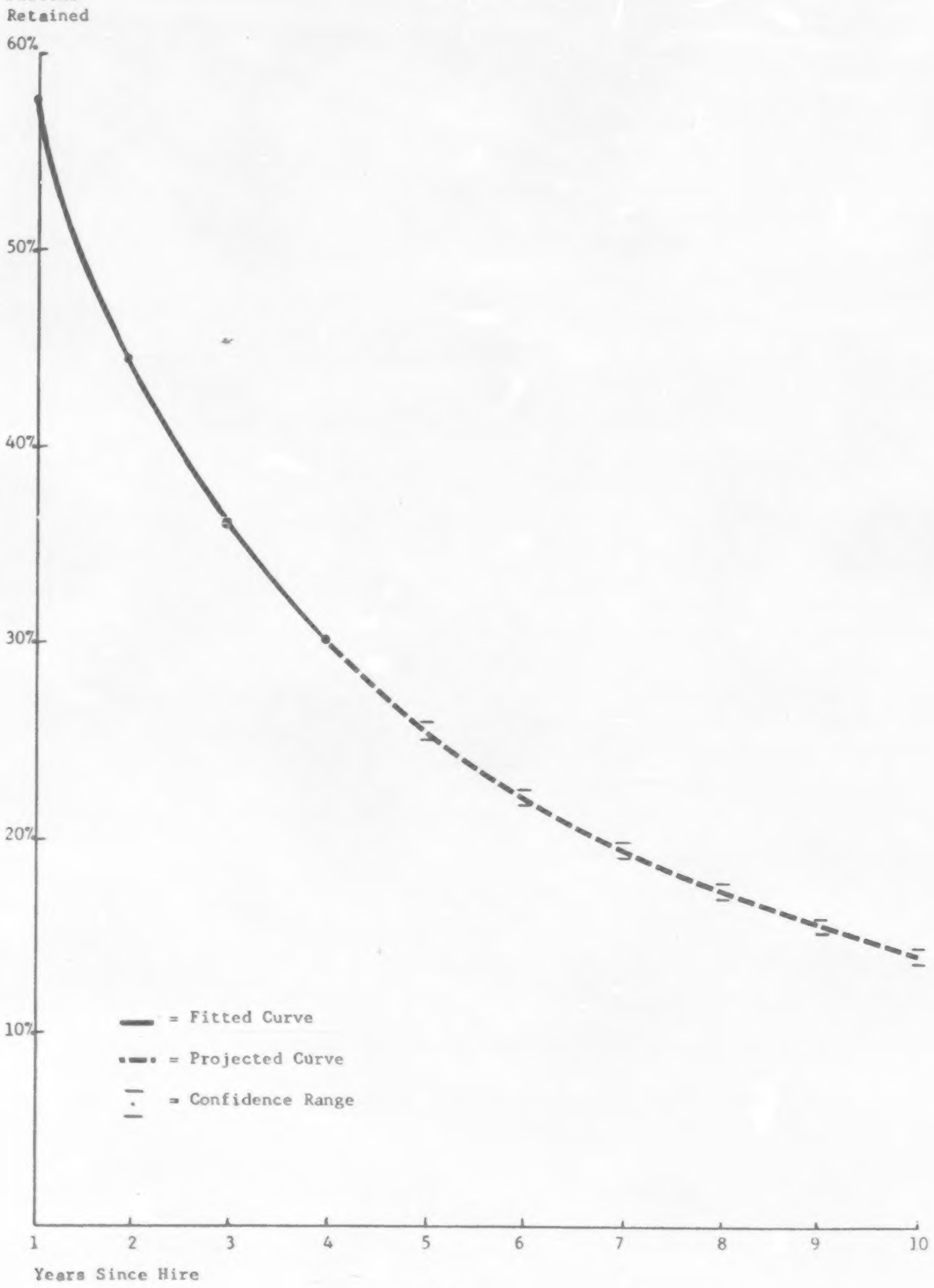
Step 6. To convert these percent figures to numbers of employees to be retained, simply multiply the number in the original group by each of the percents (in their decimal form). Suppose, for this example, that there were 250 employees in the starting group. Then the range values would be:

Years Since Hire	Range (in Numbers) Upper Limit	Range (in Numbers) Projected Value	Range (in Numbers) Lower Limit
5	64.9	63.7	62.5
6	56.3	55.2	54.1
7	49.5	48.5	47.5
8	44.0	43.0	42.1
9	39.4	38.5	37.6
10	35.6	34.7	33.9

These figures may be rounded for planning purposes.

Each of the functions involved in fitting and projecting a retention curve that has been discussed in this chapter can be done by an already developed computer program. This program is the subject of the next chapter.

Figure 5-2


Percent
Retained
60%
50%
40%
30%
20%
10%
= Fitted Curve
= Projected Curve
= Confidence Range
1
2
3
4
5
6
7
8
9
10
Years Since Hire

CHAPTER 6

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LOGPRO

Analysis Functions

The computer program LOGPRO has been designed to automatically perform the basic retention analysis functions described in the preceding chapter. These functions are:

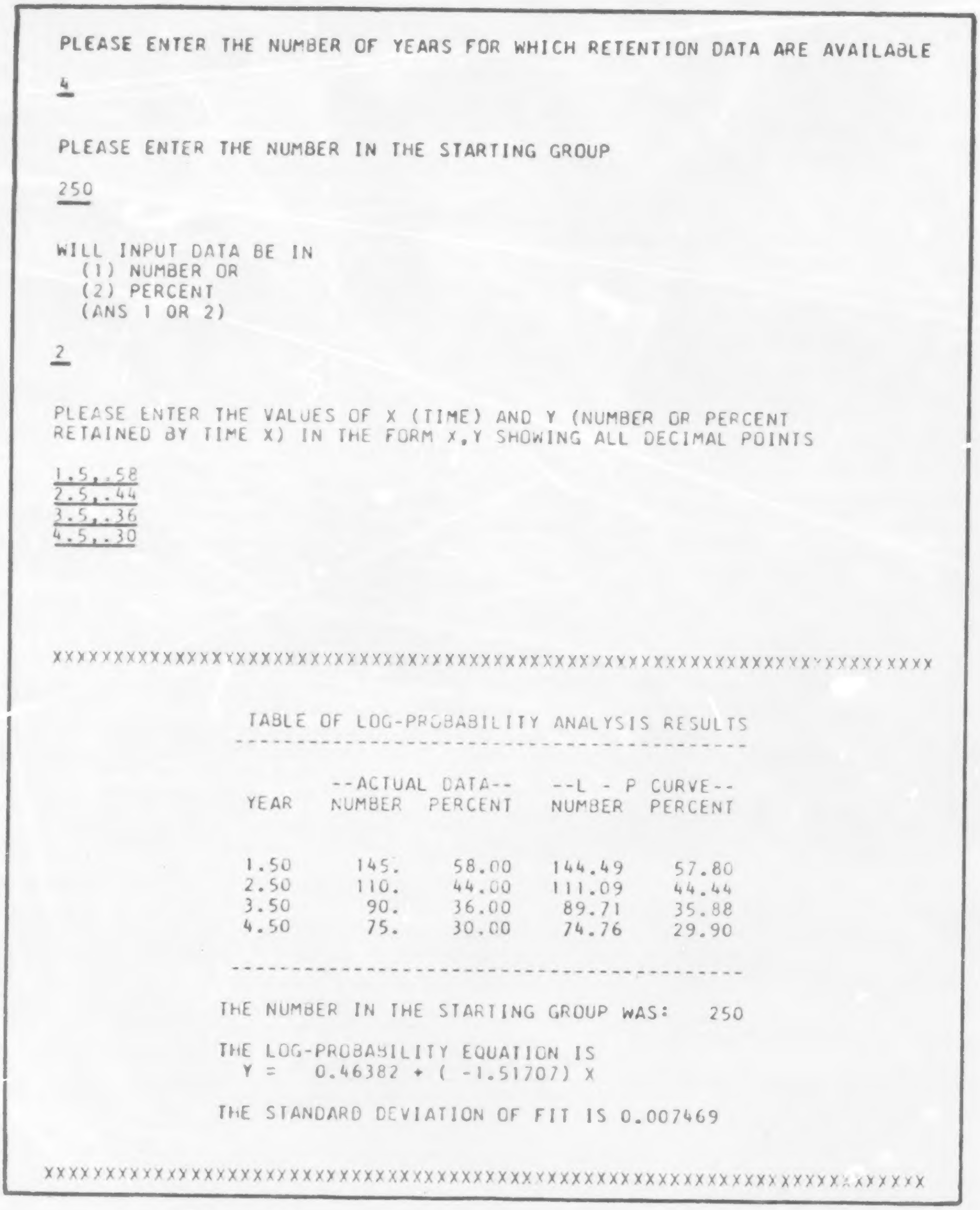
1. The calculation of the a and b values of a log-probability equation using:
   a. Operator-supplied retention data and
   b. Least squares techniques;
2. The iteration of the log-probability equation for each operator-supplied x-value;
3. The computation of the standard error about the log-probability line;
4. The projection of the log-probability equation for an operator-supplied projection period; and
5. The calculation of the 95% confidence range for each projected value.

LOGPRO (and each of the other staffing needs planning computer programs discussed in this handbook) forms a self-contained, comprehensive package consisting of (a) the main program (LOGPRO) which performs the retention analysis and (b) the subprograms (ANDPX and ANDXP) which handle the transformations from "percent retained" to "number of standard deviations from the mean of the normal curve" and back again. A user of this set of programs need not have an extensive statistical background in order to successfully utilize and evaluate its results.

LOGPRO is presently set up to be used in a time-sharing environment with the user supplying retention data from a remote terminal during the run of the program. In other words, the program will ask the user for information and the user will provide it.

A run of LOGPRO consists of two main analysis sequences. They are:

Figure 6-1



```
PLEASE ENTER THE NUMBER OF YEARS FOR WHICH RETENTION DATA ARE AVAILABLE

4


PLEASE ENTER THE NUMBER IN THE STARTING GROUP

250


WILL INPUT DATA BE IN
  (1) NUMBER OR
  (2) PERCENT
  (ANS 1 OR 2)

2


PLEASE ENTER THE VALUES OF X (TIME) AND Y (NUMBER OR PERCENT
RETAINED BY TIME X) IN THE FORM X,Y SHOWING ALL DECIMAL POINTS

1.5,.58
2.5,.44
3.5,.36
4.5,.30




XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

                TABLE OF LOG-PROBABILITY ANALYSIS RESULTS
               ------------------------------------------

                       --ACTUAL DATA--     --L - P CURVE--
               YEAR    NUMBER  PERCENT     NUMBER  PERCENT


               1.50      145.    58.00     144.49    57.80
               2.50      110.    44.00     111.09    44.44
               3.50       90.    36.00      89.71    35.88
               4.50       75.    30.00      74.76    29.90

               ------------------------------------------

              THE NUMBER IN THE STARTING GROUP WAS:   250

              THE LOG-PROBABILITY EQUATION IS
                Y =   0.46382 + ( -1.51707) X

              THE STANDARD DEVIATION OF FIT IS 0.007469

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

1. The log-probability analysis of the given retention data; and
2. The projection of the calculated log-probability equation.

Log-Probability Analysis

During this sequence, LOGPRO performs all the calculations required to fit a log-probability line to longitudinal retention data (including functions 1-3 above).

The information fed to the computer by the operator for this phase of the program consists of:

1. The number of years for which retention data are available; i.e., the number of known retention points.
2. The number of employees in the original or starting group for the occupation to be analyzed.
3. Whether the collected retention data are in the form "number of employees retained" or "percent of employees retained."
4. The actual retention data.

An example of the question and answer sequence for this part of LOGPRO is shown in Figure 6-1. Respcnses made by the operator are underscored. After each request for information, the computer pauses and waits for the operator to respond to the request. Each user response is followed by a carriage return. This sends control of the program back to the computer.

The data used in the sample run are:

Years Since Hire	Percent Retained
1	58%
2	44%
3	36%
4	30%

Number in the original group = 250
Averaging factor = 0.5 years

Since there are four pairs of retention points, the answer to the first request for information is simply 4. And, since there were 250 employees in the starting group, the answer to the second request is 250.

The next question requests that the operator make a choice based on the form of the retention data. If data have been collected in "number of employees retained" (i.e., without conversion to percent) then alternative (1) is desired and the operator response is 1. In this case, the program will automatically convert these values to percent retained.

If the data have been manually converted to percent retained, then alternative (2) is desired. And the operator response is 2. Since this is the form of our sample data, the response in Figure 6-1 is 2.

Next comes the entry of the actual retention data. One important point to remember for correct entry of these data is that all decimal points must be shown. This rule applies to every piece of data entered at this time including whole numbers. For example, 1 year is entered as 1., 100 employees is entered as 100.

In addition, when the data are in percent, the decimal form of the percent value is entered. For example, 58% is entered as .58, 58.62% is entered as .5862.

Each retention pair is entered on a single line in the form:

time-value, retention-value CR

Entry continues until the supply of retention pairs is exhausted. LOGPRO takes this retention data and converts it to log-probability form. Then the linear least squares technique is used to determine the a and b values of the log-probability equation. This equation is iterated for each given time value and the standard error is computed. The results of all these calculations are then printed out in the "Table of Log-Probability Analysis Results" (Figure 6-1).

As you can see, this table gives you, in a convenient form, both the inputted retention values ("ACTUAL DATA") and the results of the log-probability analysis ("L-P CURVE") so that they can be readily compared. In addition, the computer prints out the number in the starting group, the a and b values of the log-probability equation and the standard error of the log-probability fit (in standard deviation units).

Projection Analysis

During this sequence, LOGPRO performs all the calculations required to project a log-probability equation and calculate the 95% confidence range for each projected value.

The information given to the computer during this phase of LOGPRO consists of:

1. Whether the operator wishes to project the log-probability equation (a yes or no decision); and

2. The time parameters of the projection.

An example of the question and answer sequence for this phase of LOGPRO is found in Figure 6-2. This sample run projects the log-probability equation calculated in Figure 6-1.

At the beginning of the projection section, the operator may choose whether or not to project the calculated log-probability equation. If not, the computer will skip around this sequence. If yes, the computer will ask for certain information concerning the time frame for which the projection is wanted. At this point, the operator must enter three values. They are:

1. The first (or minimum) time value for which a projection is wanted.

2. The last (or maximum) time value for which a projection is wanted.

3. The length of the time interval between successive projection points (or the time increment).

These values are entered side-by-side, separated by commas (as shown in Figure 6-2). Here again all decimal points must be shown.

The sample run in Figure 6-2 projects the log-probability equation for the time values 5.5 years to 10.5 years. The increment in this case is 1 year. This means that projection points will be calculated for each of these values: 5.5, 6.5, 7.5, 8.5, 9.5, and 10.5.

LOGPRO now takes these time values and projects the log-probability equation for the given time frame. It also calculates the 95% confidence ranges for each projected value. These ranges are calculated in both projected numbers of employees and projected percent retained. The results of all of these calculations are printed out in the "Table of Projected Values." This table contains, for each time value, the projection (EXPECTED VALUE) and its 95% range for both projected numbers and projected percents.

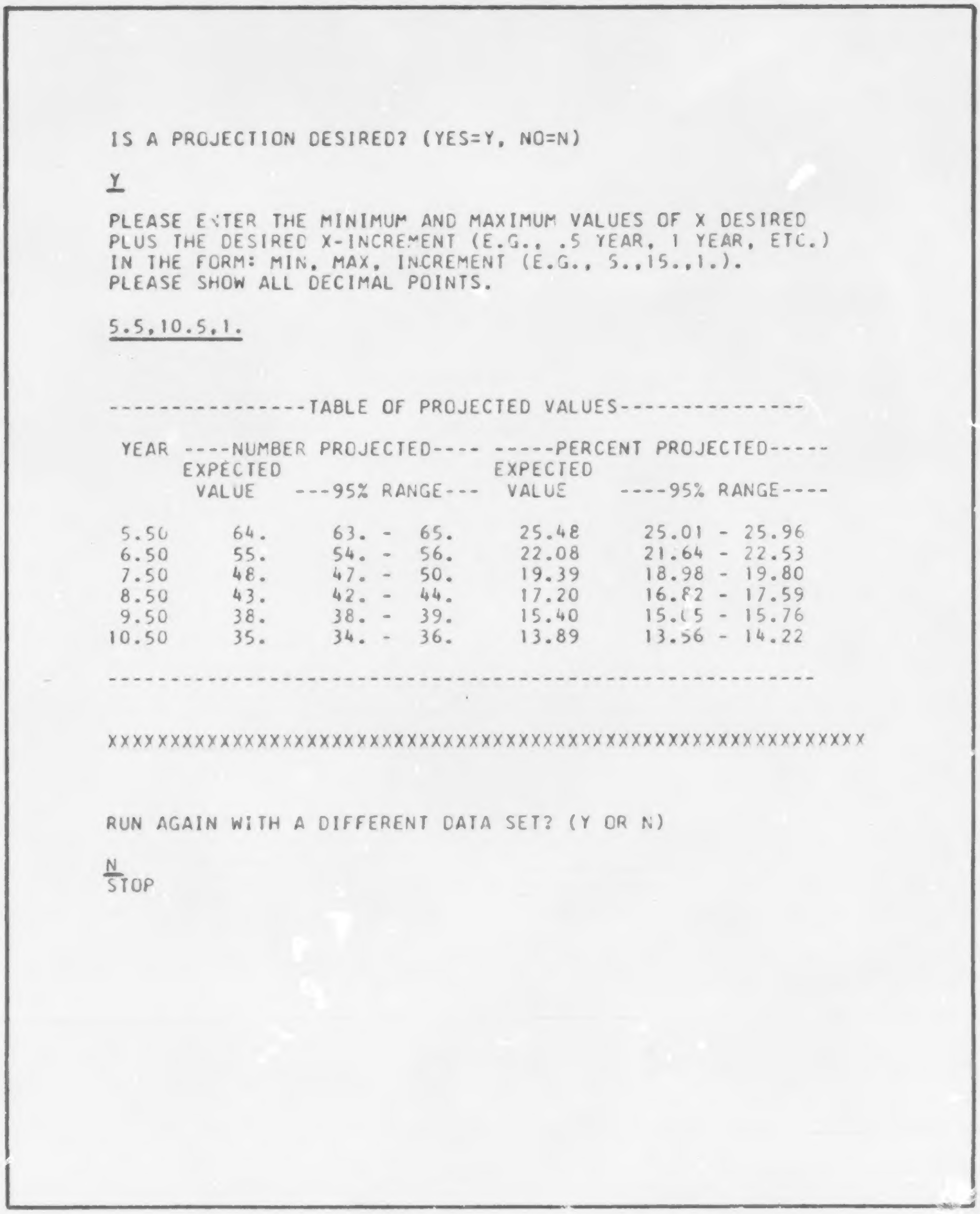
After printing out the projection table, the computer will ask the operator if there are any more data to be analyzed. If there are, the computer will recycle to the beginning of LOGPRO. If not, the computer will stop the execution of the program.

Summary

LOGPRO and its subprograms form a complete statistical package for the analysis of retention data which are:

85

Figure 6-2



```
IS A PROJECTION DESIRED? (YES=Y, NO=N)

Y

PLEASE ENTER THE MINIMUM AND MAXIMUM VALUES OF X DESIRED
PLUS THE DESIRED X-INCREMENT (E.G., .5 YEAR, 1 YEAR, ETC.)
IN THE FORM: MIN, MAX, INCREMENT (E.G., 5.,15.,1.).
PLEASE SHOW ALL DECIMAL POINTS.

5.5,10.5,1.


----------------TABLE OF PROJECTED VALUES---------------

 YEAR ----NUMBER PROJECTED---- -----PERCENT PROJECTED-----
      EXPECTED                 EXPECTED
       VALUE   ---95% RANGE---  VALUE     ----95% RANGE----

 5.50    64.      63. -  65.     25.48     25.01 - 25.96
 6.50    55.      54. -  56.     22.08     21.64 - 22.53
 7.50    48.      47. -  50.     19.39     18.98 - 19.80
 8.50    43.      42. -  44.     17.20     16.82 - 17.59
 9.50    38.      38. -  39.     15.40     15.05 - 15.76
10.50    35.      34. -  36.     13.89     13.56 - 14.22

----------------------------------------------------------


XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX


RUN AGAIN WITH A DIFFERENT DATA SET? (Y OR N)

N
STOP
```

1. Longitudinal - i.e., employee groups (or cohorts) are followed over time and counts are made of how many employees remain on board after certain lengths of time.

2. Occupational - i.e., data are collected by occupation in all cases. Data are also entered by occupation except in those cases where aggregation of occupations is possible (see Chapter 8).

LOGPRO uses the statistical techniques described in Chapters 4 and 5 to perform its analyses. A more detailed discussion of the mathematics used by LOGPRO can be found in its Technical Analysis which can be found in the Appendix section along with its Operation Manual and Program Listing.

Although LOGPRO is an extremely useful tool for retention analysis, there are some situations where it is infeasible to use it directly, such as in organizations with:

- Hiring patterns which vary widely from year to year.
- Low levels of hiring.
- Many occupations having only a few employees.

The techniques used to handle these and similar situations will be explained in the next two chapters.

CHAPTER 7

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LPFILE

So far we have been discussing techniques of retention analysis which require the use of longitudinal data. However, there are cases where either (a) longitudinal data are not available or easily obtainable or (b) organizational hiring patterns do not readily adapt to a LOGPRO type of analysis. We now need a way to apply the log-probability analysis technology to such cases. For this purpose we have developed a second retention analysis computer program. This second program is known as LPFILE.

When To Use LPFILE

There are certain situations where longitudinal retention analysis becomes less precise and, thus, where the use of LPFILE is recommended. The three most common of these situations will be discussed.

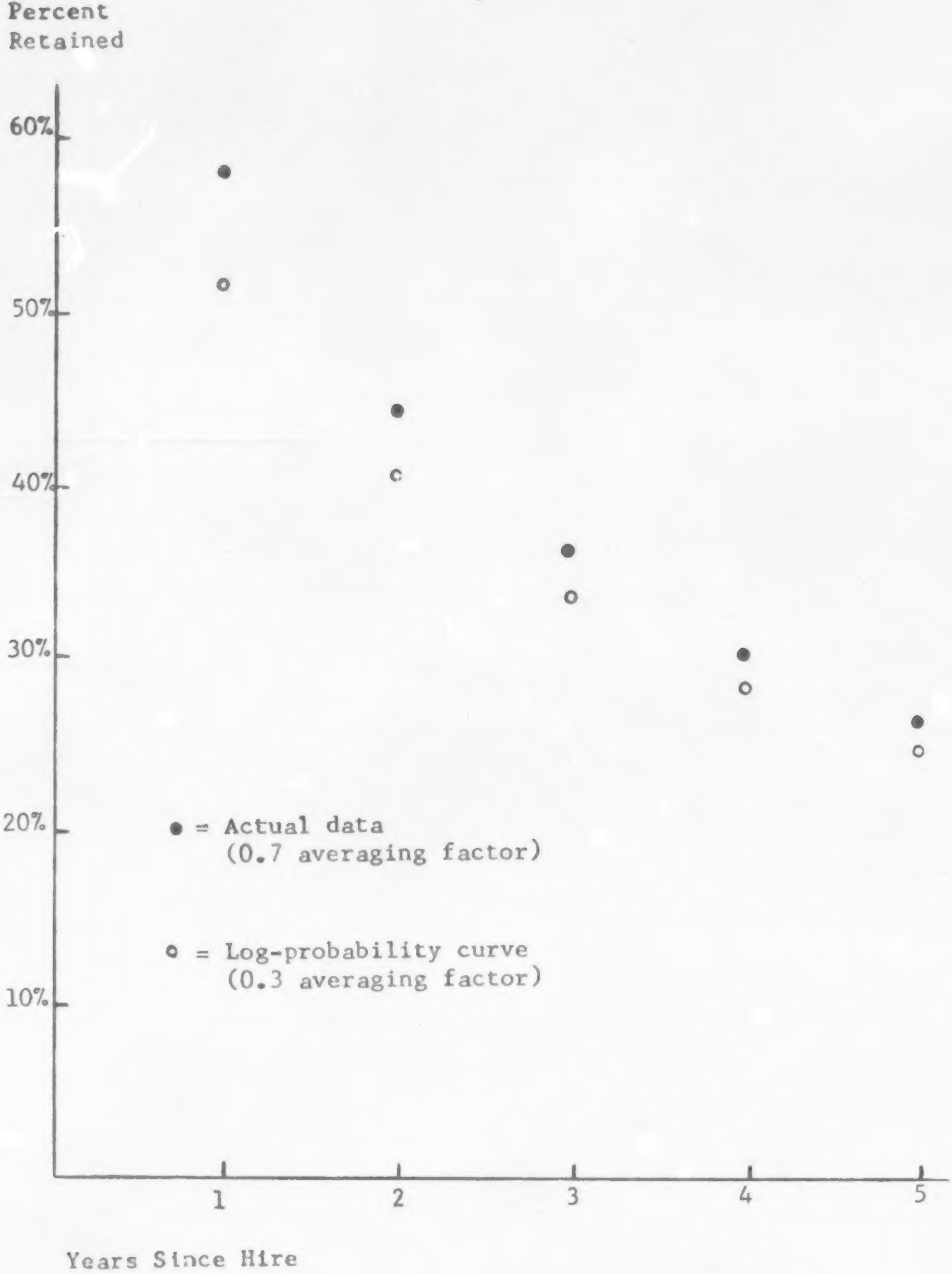
1. The number of employees hired into one or more of an organization's occupations during any one given year is relatively small.

This type of situation can occur both in small organizations with few employees and in large organizations consisting of a scattering of many occupations with few employees. In such cases, the use of longitudinal data is limited because a low level of hiring leads to small cohort groups. And small cohort groups, like small statistical samples, are more subject to error than are large groups. Also, a small cohort is less likely to have a truly normal distribution of work attitudes (since the likelihood of the existence of a normal curve increases as the number of work attitudes increases). In addition, for a small group, the relationships between standard deviation values and percentage of cases change considerably.

2. The yearly pattern of hiring into an occupation changes from one year to the next.

One example of this type of situation occurs when hiring into a given occupation is concentrated early in one year and late in another. Such yearly differences in hiring patterns will cause yearly changes in the time-averaging factor which is used to calculate the log-probability curve (e.g., 0.3 in one year and 0.7 in the next). This means that a new equation must be computed for each year's group of hires which in turn means continually collecting and recording longitudinal data for each group. Thus, such a situation makes it difficult to apply a log-probability equation computed from a group of employees hired in one given year to employees hired in any other year.

Figure 7-1


Percent Retained
60%
50%
40%
30%
20%
10%
1
2
3
4
5
Years Since Hire
● = Actual data (0.7 averaging factor)
○ = Log-probability curve (0.3 averaging factor)

For example, Figure 7-1 plots a log-probability equation computed using an averaging factor of 0.3 years against longitudinal data for a group of hires for which the actual averaging factor was 0.7 years. As you can see, any type of projection or calculation using this equation for this particular group of employees would be consistently low.

In addition, under these conditions, the use of a single log-probability equation to perform the analysis applications discussed in Chapter 5 becomes imprecise. This is because the applications were designed to be used on all employees in a given occupation regardless of their year of hire.

As a corollary to this situation, it might be that hiring into an occupation during a given year is so unevenly distributed as to make it difficult to use a convenient time-averaging factor. In such cases the only recourse would be to compute an actual average-length-of-service figure using the entry-on-duty date of every employee in the cohort. This may be too cumbersome a task, particularly if the resultant averaging factor cannot be applied to any other year's data.

3. One year's log-probability retention curve for a given occupation is noticeably different from another year's curve for the same occupation.

It may be that observation of different years' cohort groups for an occupation reveals noticeable (although not statistically significant) differences in the retention curve for the occupation. This usually happens when the number of employees hired in the individual years is relatively small, so that random variation differences tend to be relatively large. What is needed in such cases is a way to average the curve differences--which is what LPFILE will do.

The LPFILE Method

The basic motivating factor behind the development of LPFILE was the need for a method which would handle each of the above situations by making the maximum possible use of every bit of information available. Data on every employee hired into an occupation during a selected time span (say, 5 years) are inputted to LPFILE. These data are used to compute a log-probability equation for the occupation. The system used has several advantages:

- Since data on every employee are utilized, every available contributor to the group retention trend is used to compute the group's equation.

- The size of the sample used to calculate the equation is maximized. (This is particularly important when hiring levels are low.)

- Use of all the years of data available gives the maximum possible precision to the result.

- The employees used to compute the equation are not limited to one year's hires.

- There is no need to continually collect and record longitudinal data.

LPFILE uses a previously-stored file of employee data to compute a log-probability retention equation. In this section we will discuss first the mathematics used by LPFILE, then the set-up of the data file, and lastly examples for specific sets of data.

The Mathematics: Suppose that we knew for the past, say, five years (a) the date when every employee was hired into a given occupation and (b) the date when those employees who left the occupation did so. With this knowledge we could develop a quasi-longitudinal retention curve for the occupation to which we could apply the log-probability analysis technique.

We know from the previous discussions of log-probability analysis that two items of data are needed for use of the linear least squares technique: (1) the percent retained from a given group of employees at (2) specific times. For LPFILE we need to collect our data so that we can construct replicas of these data items. To do this we first need to select a time period for which data collection will take place. This could be the past two years, three years or whatever time span for which your organization keeps chronological records of employees hired.

Second, we record, for every employee hired into a given occupation during that time period, one or two dates:

1. The date of hire into the occupation (for every employee); and

2. The date of separation from the occupation (for those employees who have separated[1] during the time span).

These dates become part of the data file which is used by LPFILE.

We need one other date for the use of the LPFILE method and it is known as the "file ending date." This is the closing date of the file and of the time period under study.

---

[1] Note that, as discussed at the beginning of Chapter 4, a "separation" is any personnel action which creates a vacancy in the group being analyzed. Thus, a movement from this to some other occupation within the same organization is a "separation."

LPFILE uses these dates to calculate two numbers for every employee in the group. The values of these numbers will depend on whether or not an employee has separated from the group.

If an employee has separated, the numbers computed are:

1. The employee's actual time on board in the occupation (i.e., Date of Separation - Date of Hire); and

2. The length of time that the employee could possibly have been on board during the selected time span if he or she had not separated before the end of that time (i.e., File Ending Date - Date of Hire).

If an employee has not separated, his or her actual time on board will be equal to his or her possible time on board. So, for these employees the two computed values are equal.

LPFILE records each of the computed values in one of two arrays. The first array consists of the values of the actual time on board for every employee in the group. The second consists of the values of the elapsed time between every employee's date of hire and the file ending date. (These last values are either "possible time on board" or "actual time on board" depending on whether or not an employee has separated.)

These computed values are then used to develop a distribution comparing the number of employees in the group who could possibly have served for a given length of time with the number of employees in the group who actually did serve for at least that same length of time.

LPFILE now uses the results of its computations to set up the **x, y pairs** for log-probability retention analysis. First, each different value from the elapsed time array becomes a length of time from original hire for those employees who could possibly have served (or who actually did serve) for that length of time. These values then become the time -- or x -- values for log-probability analysis.

The percent retained (or y) values are calculated using the following data substitutions for each x - value:

1. The number of employees retained at time x = the number of employees who actually stayed with the group for at least time x (including those employees who separated at some time later than x).

2. The number of employees in the original group of hires = the number of employees who could have stayed with the group for at least time x (including both employees who left before time x and employees who stayed longer than time x).

Figure 7-2

LPFILE DATA COLLECTION FORM

EMPLOYEE NUMBER	OCCUPATION CODE	GRADE AT HIRE	DATE OF HIRE MO. YEAR	DATE OF SEPARATION MO. YEAR

For each x - value, data item #1 is divided by data item #2 to obtain an estimate of the percent retained at that time. And these become the y-values.

LPFILE uses these x, y pairs and the data transformations found in LOGPRO to calculate a least squares log-probability equation for the given group. The resulting equation can be used in the same way as an equation derived by LOGPRO.

The Data File: To perform its calculations LPFILE needs to be supplied with a specially set up file of employee data. This file consists of one record for each employee in a group (or groups -- see discussion of possible file combinations below). There are four data elements in each record. They are the employee's:

1. Occupation or series code;
2. Grade at hire;
3. Date of hire (month and year); and
4. Date of separation (month and year), if any.

The last two items are, of course, those we discussed in connection with the method used by LPFILE. The first two items permit the user to both (a) include more than one occupation in a single file and (b) perform log-probability analysis on employees in different entry-grade levels.

In most organizations, these data items can typically be found in a chronological file of personnel actions in the personnel office. Or they may be recorded in an organization's automated information system.

Before data collection is begun, decisions must be made on (a) the time span for which data will be collected and (b) the occupations which will be studied. This information is then passed on to a computer programmer (if the data system is automated) or to the person who will collect the data manually.

If data are to be collected manually, it is helpful to use some type of a standard collection form such as the one in Figure 7-2. You will notice that an extra data item has been included in this form. This is an employee identification number (e.g., Social Security Account Number). This item is needed since the other four data elements will most likely be found on separate sheets of paper. However, LPFILE does not need it to do its work. A partial sample of a filled out collection form can be found in Figure 7-3(A).

To get the data from the collection form into a data file in the computer so that LPFILE can use it, a few rules must be followed:

Figure 7-3

A. LPFILE DATA COLLECTION FORM

EMPLOYEE NUMBER	OCCUPATION CODE	GRADE AT HIRE	DATE OF HIRE MO. YEAR	DATE OF SEPARATION MO. YEAR
1	201	5	1-73	11-74
2	201	5	11-72	NONE
3	201	5	7-72	1-75
4	201	5	9-71	3-75
5	201	5	7-73	NONE
6	201	5	9-73	5-75
7	201	5	4-72	NONE
8	201	5	1-72	2-75
9	201	5	5-74	NONE
10	201	5	11-73	6-74

B. COMPUTER DATA FILE

```
002010501731174
002010511720000
002010507720175
002010509710375
002010507730000
002010509730575
002010504720000
002010501720275
002010505740000
002010511730674
```

FILE ENDING DATE = JUNE 1975

C. LPFILE TIME VALUES

EMPLOYEE NUMBER	COULD HAVE SERVED (YEARS)	ACTUALLY SERVED (YEARS)
1	2.42	1.83
2	2.58	2.58
3	2.92	2.50
4	3.75	3.50
5	1.92	1.92
6	1.75	1.67
7	3.17	3.17
8	3.42	3.08
9	1.08	1.08
10	1.58	0.58

1. Each line on the collection form becomes a line of the data file.

2. The data elements in each line are placed in a specific order with occupation code first, then grade at hire, month of hire, year of hire, month of separation, and year of separation.

3. If an employee has not separated then the entries for both the month and year of separation are zero.

4. Depending on the available computer system, the data elements in each line are either (a) separated by commas or (b) spaced according to the criteria set up by LPFILE.

For rule 4b, the following spaces are allotted for each data element:

- Occupation code--5 spaces. (To handle series codes of up to 5 digits).

- Grade at hire--2 spaces.

- Month and year of hire, month and year of separation--2 spaces each.

If a data element requires less than the allotted space, then the rest of the area designated for that element is filled with leading zeroes or blanks.

As suggested in rule 4, above, there are two ways in which the data from an LPFILE collection form may be transformed into a computer-acceptable file. First, if the data elements in a line are to be separated by commas, then the first two rows from the sample collection form in Figure 7-3(A) would translate to:

```
201, 5, 1, 73, 11, 74,
201, 5, 11, 72, 0, 0
```

Note that, since the second employee has not separated, the last two values in the second line are zero.

Second, if the data elements in a line are to be spaced according to LPFILE's specifications, then the first two lines of the sample collection form would be entered as:

```
002010501731174
002010511720000
```

Note that (a) the elements follow each other without any separation character between them and (b) leading zeroes are used to fill unused spaces. (In most systems blanks may be used in place of the leading zeroes.) The completed file based on the sample collection form may be found in Figure 7-3(B).

The transfer of data from a collection form into the computer may be done directly or there may be an intermediate step which transfers data first to a coding sheet (where the data are arranged in computer-acceptable form) then to the computer.

In any case, entry of the file into the computer from the keyboard of a computer terminal follows standard steps which include:

1. Giving the data file its own name;
2. Entering the data file one line at a time;
3. Hitting the "carriage return" key at the end of each line; and
4. Storing the file in the computer's memory area.

Once entered and stored, a data file can be recalled at any time for use by LPFILE.

The contents of these data files may consist of one occupation per file or multiple occupations per file. A single file may also contain more than one grade at hire level. During a run of LPFILE, the user may select what portions of the file are to be analyzed. For example, if a file contains records on three occupations only one of which is to be analyzed during a given run, then LPFILE will use only those records of employees in that occupation.

Each file will have a file ending date which will be entered from the terminal keyboard during a run of LPFILE.

Examples: As a simple example of the techniques used by LPFILE, we will use the sample file of 10 employees found in Figure 7-3(A). Of course, it is unlikely that such a small file would exhibit log-normal characteristics and an actual file would consist of many more employees. But a small file is useful for illustration purposes.

Let's assume that the file ending date for this sample file is June 1975.

First we have to calculate, for each employee in the file the values for actual and possible time on board. For example, consider the data for employee #1 in Figure 7-3(A).

Date of hire = 1/73
Date of separation = 11/74

This employee could have served from January 1973 to June 1975 (the file ending date). In other words, the employee could have served for 2 years and 5 months or 2.42 years. But the employee actually served until November 1974 or 1 year and 10 months or 1.83 years.

Each employee's time values are calculated in the same manner. And for those employees who have not separated, the two values are equal. For example, for employee #2, the actual and possible values are both 2.58 years (from November 1972 to June 1975).

The completed table of actual and possible time values is found in Figure 7-3(C). Arranging these values from lowest to highest possible service gives:

Could Have Served (Years)	Actually Served (Years)
1.08	1.08
1.58	0.58
1.75	1.67
1.92	1.92
2.42	1.83
2.58	2.58
2.92	2.50
3.17	3.17
3.42	3.08
3.75	3.50

The next step is to translate these values into log-probability x, y retention pairs. The "Could Have Served" column becomes the x-values. These represent "Years Since Hire" values. To determine the percent retained for each x-value two counts are made:

1. The number of employees in the group who could have served to time x; and
2. The number of employees in the group who did serve to at least time x.

For example, the first x-value is 1.08 years. All ten employees in the group could have served for this time but only nine of them actually did (employee #10 left after 0.58 years of service). This gives a retention ratio of 9/10 or 90%.

Figure 7-4

```
008010505700475        008010509740000        008010702740000
008010506700000        008010510741174        008010703740000
008010507700175        008010511740175        008010704740874
008010508700000        008010512740000        008010705740000
008010509701271        008010501750375        008010706740000
008010510700000        008010502750000        008010707740974
008010510700000        008010503750000        008010708740000
008010511700000        008010504750000        008010709740000
008010512700000        008010505750000        008010710740000
008010501710175        008010705700974        008010711740000
008010502710000        008010706700000        008010712740000
008010503710000        008010707701274        008010701750000
008010504711072        008010708700000        008010702750000
008010505710000        008010709700473        008010703750000
008010506710000        008010710700000        008010704750000
008010507710275        008010711700000        008010705750000
008010508710000        008010712700000
008010509710000        008010701710275
008010510711274        008010702710000
008010511710000        008010703710874
008010512710000        008010704710000
008010501721174        008010705711172
008010502720000        008010706710375
008010503720974        008010707710000
008010504720000        008010708711272
008010505721272        008010709710000
008010506720274        008010710711174
008010507720000        008010711710000
008010508720000        008010712710000
008010509720000        008010701721273
008010510721274        008010702721274
008010511720000        008010703720000
008010511720000        008010704720000
008010512720000        008010705720174
008010501731274        008010706721074
008010502730000        008010707720000
008010503731174        008010708720000
008010504730000        008010709720000
008010505730000        008010710721174
008010506730000        008010711720000
008010507730874        008010712720000
008010508730000        008010701730275
008010509730000        008010702730274
008010510730374        008010703730000
008010511730000        008010704730275
008010512730000        008010705730000
008010501741274        008010706730274
008010502740000        008010707730000
008010503740000        008010708730000
008010504740275        008010709730000
008010505740000        008010710730000
008010506740974        008010711730674
008010507740000        008010712730000
008010508741274        008010701740674
```

The next x-value is 1.58 years. Nine of the employees in the group could have served for this time (employee #9 is no longer in the running). Of these nine, eight actually did serve for at least 1.58 years. Thus the retention ratio for 1.58 years is 8/9 or 89%. And so on through each x-value.

The retention values for this sample file would be:

Years Since Hire (x)	Retention Ratio	Percent Retained (y)
1.08	9/10	90
1.58	8/9	89
1.75	7/8	88
1.92	6/7	86
2.42	5/6	83
2.58	4/5	80
2.92	3/4	75
3.17	2/3	67
3.42	1/2	50
(3.75	0/1	0)

(Since there was no one left to contribute to the group retention trend by 3.75 years, this time value would not enter into the log-probability equation calculations.)

LPFILE utilizes the linear least squares technique on these x- and y-values to compute a log-probability equation for the given group.

As we said before, a file consisting of only ten employees is unlikely to exhibit true log-normal behavior. To illustrate what LPFILE data will look like in a practical situation, we have constructed two larger sample files. The first is a hypothetical file for a professional occupation. For illustration purposes, we have used the GS-801 (General Engineering) occupation. This file, which we have named LP801, is listed in Figure 7-4. The second sample file represents a clerical occupation. For this file, we have used GS-322 (Clerk-Typist). We have named it LP322 (see Figure 7-5).

To show the type of retention pattern which is exhibited by such files, we have plotted in Figure 7-6 the retention pairs derived from file LP322 (the scattered dots). As you can see, there is a distinct and fairly even log-normal retention pattern exhibited by these data. The pattern deviates somewhat in the tail area of the curve. This is due to the smaller numbers of employees who contribute to the trend at the higher length-of-service levels.

The log-probability curve which LPFILE fitted to the LP322 data is shown by the dotted line in Figure 7-6.

Figure 7-5

```
003220301701173
003220302700874
003220303701074
003220304700000
003220305700774
003220306700273
003220307700973
003220308700173
003220309700873
003220310700772
003220311700471
003220312700173
003220301710473
003220302710173
003220303710000
003220304711172
003220305710000
003220306711272
003220307710972
003220308710172
003220309711072
003220310710272
003220311710000
003220312710473
003220301720000
003220302720573
003220303720572
003220304720473
003220305720473
003220306720000
003220307720373
003220308720000
003220309720000
003220310720873
003220311720000
003220312720673
003220301730000
003220302731173
003220303730000
003220304730473
003220305731173
003220306730000
003220307731173
003220308730873
```

```
003220309730574
003220310731173
003220311730000
003220312730174
003220301740474
003220302740374
003220303741274
003220304741174
003220305740774
003220306740000
003220307741074
003220308740000
003220309741174
003220310741074
003220311741274
003220312740000
```

Figure 7-6

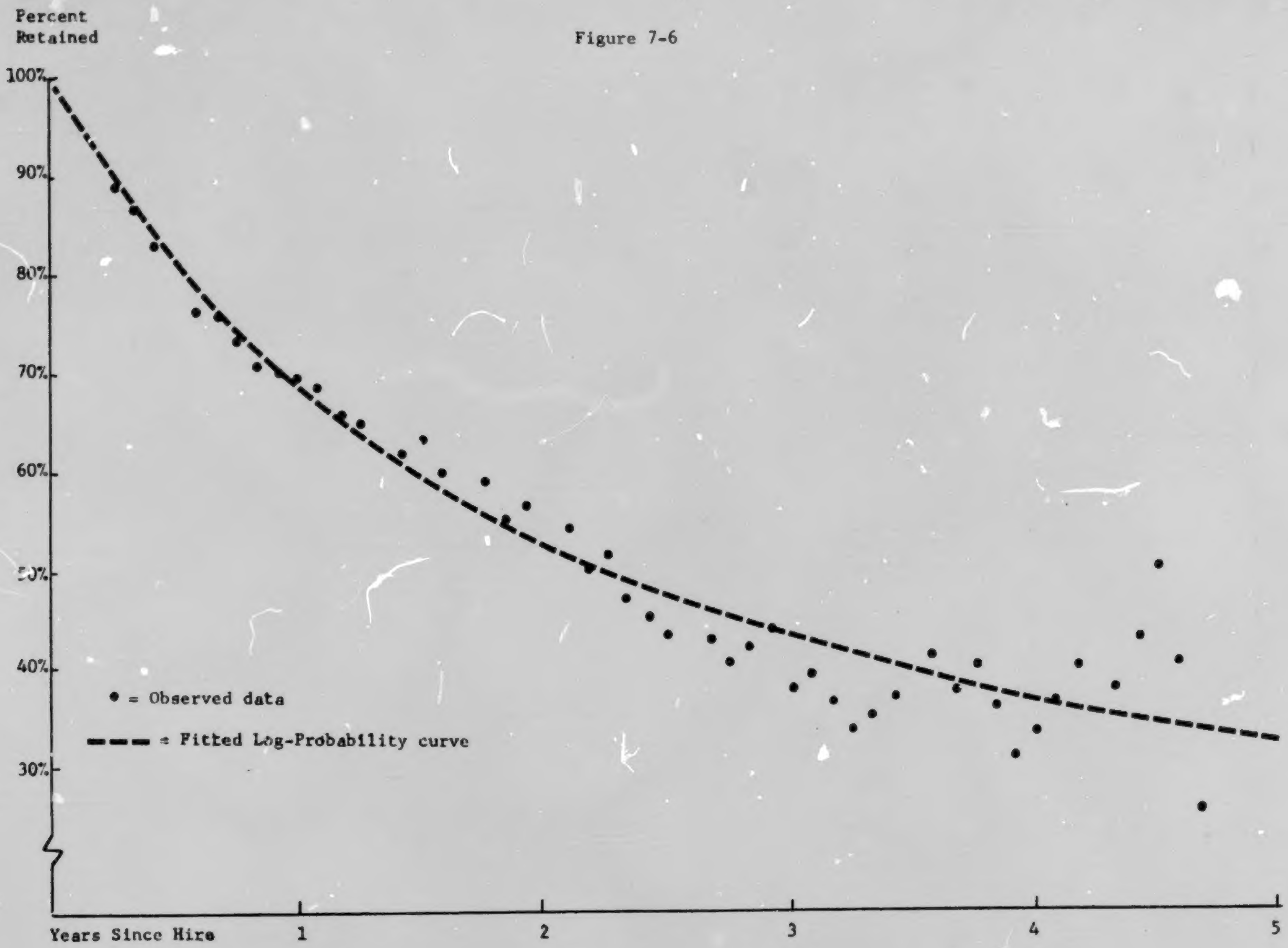
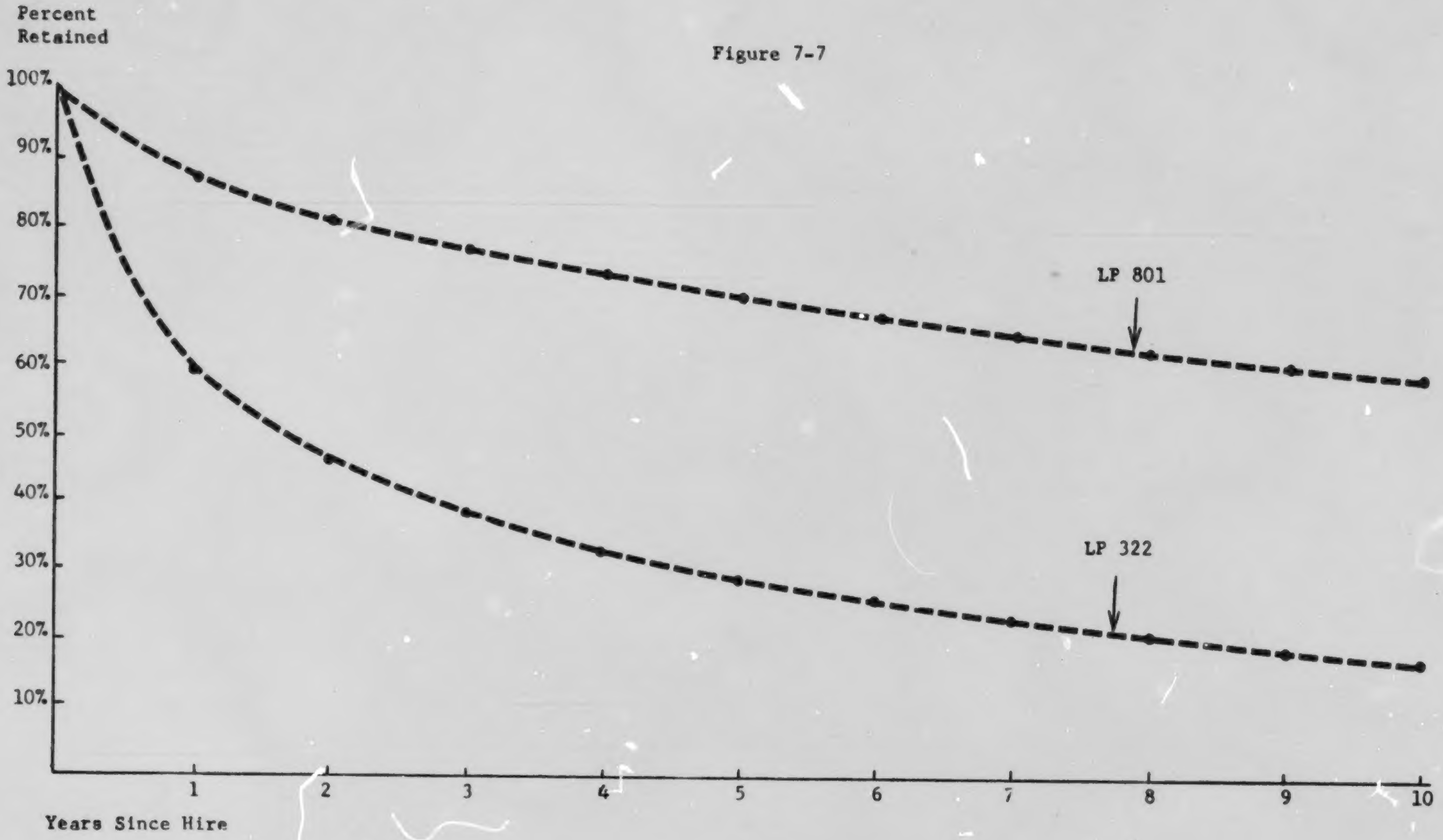

Percent Retained
100%
90%
80%
70%
60%
50%
40%
30%
● = Observed data
▬ ▬ ▬ = Fitted Log-Probability curve
Years Since Hire
1
2
3
4
5

Figure 7-7


Percent Retained
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
LP 801
LP 322
1
2
3
4
5
6
7
8
9
10
Years Since Hire

In Figure 7-7, the LPFILE fitted curves for both LP322 and LP801 are plotted. This graph shows that the existence of occupational differences as discussed for longitudinal data are still in effect under the LPFILE system. In addition, since the LP801 file contains records for both GS-5 and GS-7 hires, the two retention curves based on these different entry-levels are plotted in Figure 7-8. This graph suggests that there may be curve differences between different entry-grades that require the use of more than one log-probability equation. There will be more on such differences in the next chapter.

Sample Run of LPFILE

Figure 7-9 shows a sample run of the program LPFILE. There are two sections to this run: information inputs and analysis outputs.

The first piece of information to be input is the file ending date for the file to be analyzed. This date is entered by typing the value of the month (from 1 to 12), and the last two digits of the year (e.g., 75 for 1975), separated by commas.

The next item needed by LPFILE is the number of employees in the whole file. Or, in other words, the number of lines in the file. Then the name of the file is requested. As presently set up in LPFILE, this name can be at most five characters long. The first character must be alphabetic although numbers may be used in the other positions (e.g., LP801). This name is the same one under which the data file was stored earlier.

That portion of the file which is to be analyzed during the current run is chosen by the user's response to the next question:

DO YOU WISH BREAKDOWN BY:

(1) OCCUPATION

If a file contains records for employees in more than one occupation, these occupations should be analyzed separately by choosing this option. LPFILE will ask for the desired occupation code and select for analysis only the records of employees in that occupation. The other occupations may be analyzed by re-running LPFILE until all the occupations have been completed.

(2) GRADE

If you feel that there may be differences in the turnover curves for different entry-grade levels (although the effect of grade at hire on retention is much less significant than the effect of length of service), this option may be used. If there is more than one occupation in a file then this option alone will analyze all employees in the selected entry grade regardless of occupation.

Figure 7-8

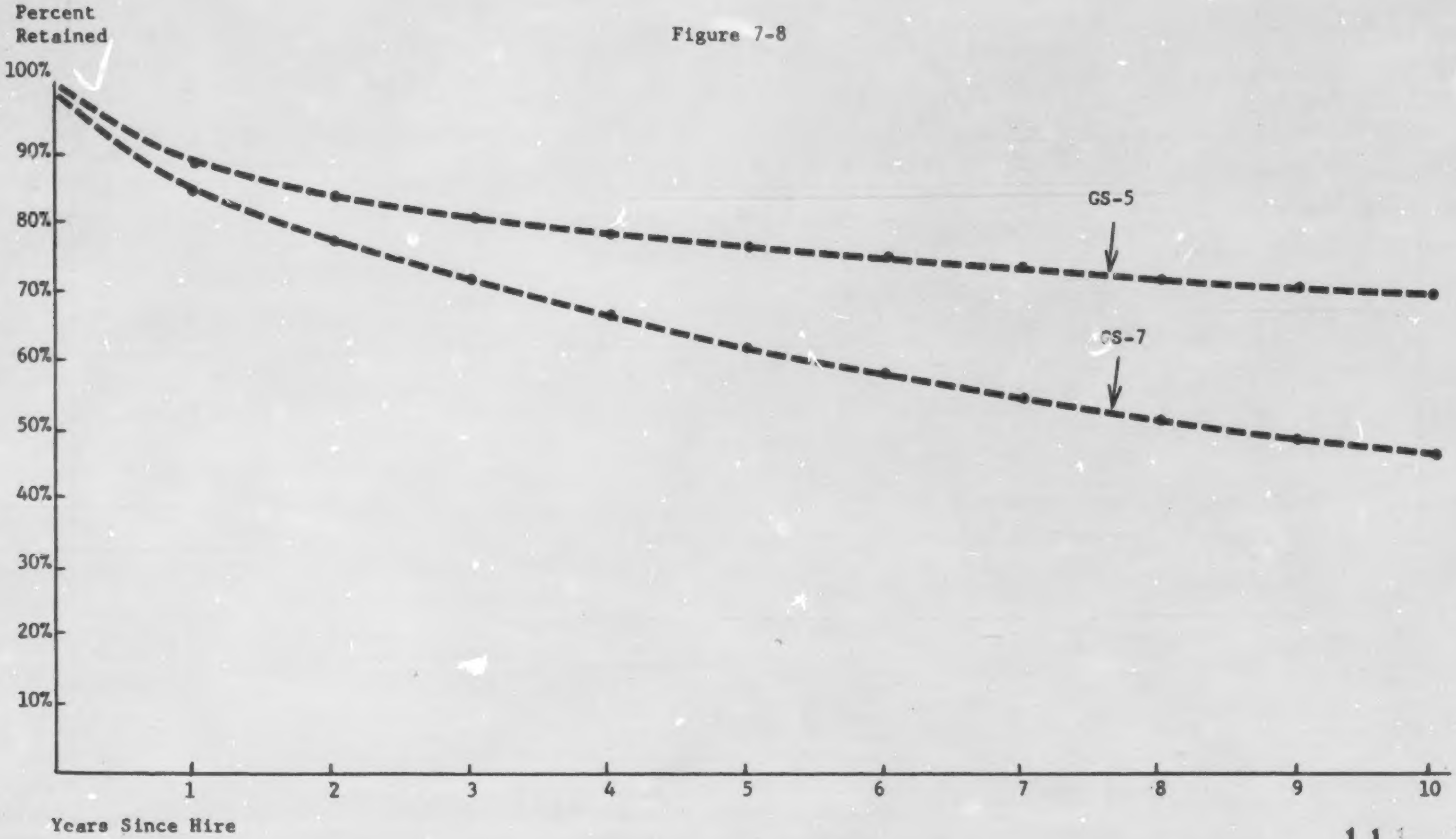
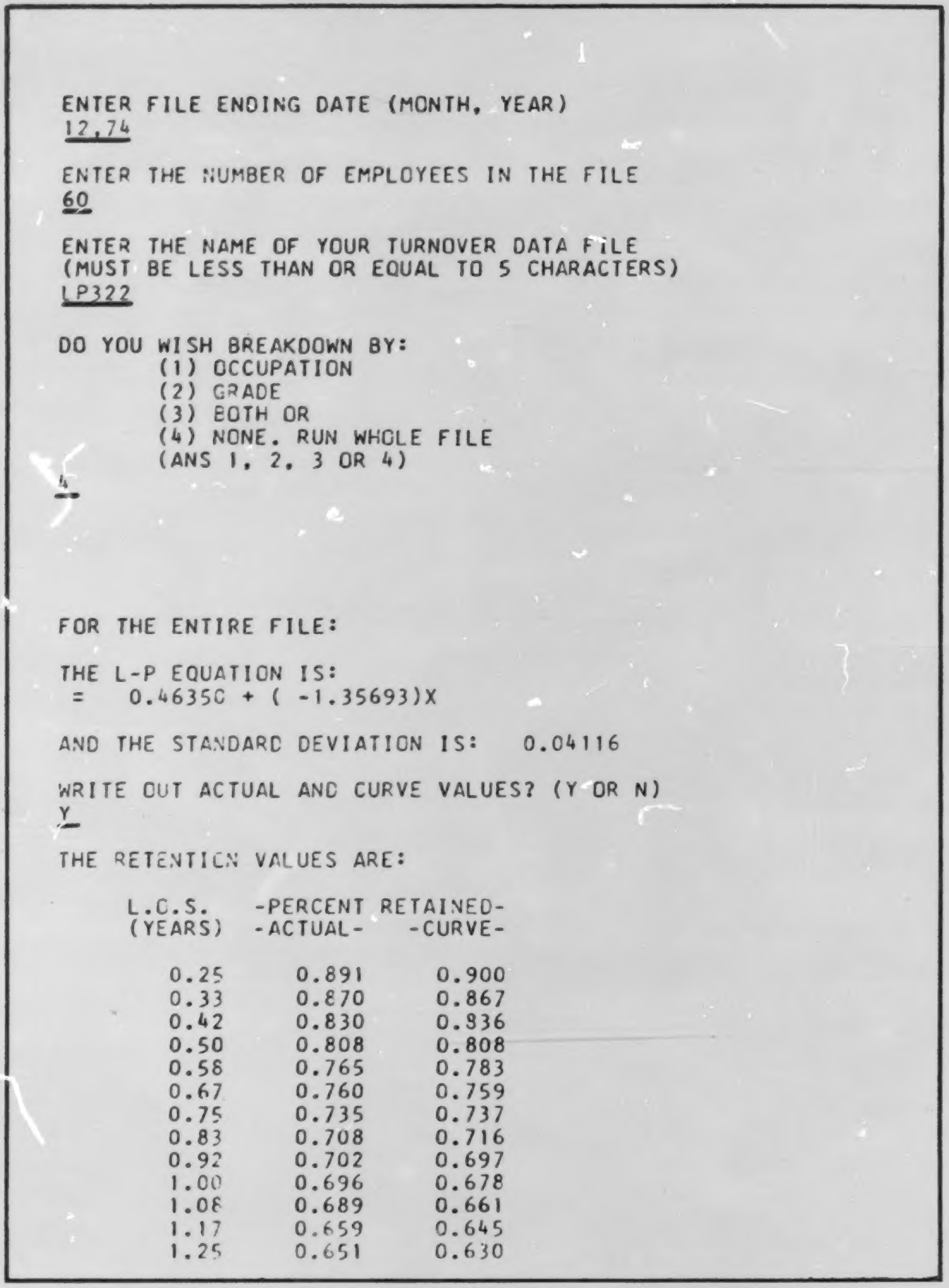

Percent Retained
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
GS-5
GS-7
1
2
3
4
5
6
7
8
9
10
Years Since Hire

Figure 7-9



```
ENTER FILE ENDING DATE (MONTH, YEAR)
12,74

ENTER THE NUMBER OF EMPLOYEES IN THE FILE
60

ENTER THE NAME OF YOUR TURNOVER DATA FILE
(MUST BE LESS THAN OR EQUAL TO 5 CHARACTERS)
LP322

DO YOU WISH BREAKDOWN BY:
       (1) OCCUPATION
       (2) GRADE
       (3) BOTH OR
       (4) NONE. RUN WHOLE FILE
       (ANS 1, 2, 3 OR 4)
4




FOR THE ENTIRE FILE:

THE L-P EQUATION IS:
 =   0.46350 + ( -1.35693)X

AND THE STANDARD DEVIATION IS:   0.04116

WRITE OUT ACTUAL AND CURVE VALUES? (Y OR N)
Y

THE RETENTION VALUES ARE:

     L.C.S.   -PERCENT RETAINED-
     (YEARS)  -ACTUAL-   -CURVE-

        0.25     0.891      0.900
        0.33     0.870      0.867
        0.42     0.830      0.836
        0.50     0.808      0.808
        0.58     0.765      0.783
        0.67     0.760      0.759
        0.75     0.735      0.737
        0.83     0.708      0.716
        0.92     0.702      0.697
        1.00     0.696      0.678
        1.08     0.689      0.661
        1.17     0.659      0.645
        1.25     0.651      0.630
```

Figure 7-9 (cont'd)

```
        1.42        0.619        0.602
        1.50        0.634        0.589
        1.58        0.600        0.576
        1.75        0.590        0.553
        1.83        0.553        0.542
        1.92        0.568        0.532
        2.00        0.528        0.522
        2.08        0.543        0.512
        2.17        0.500        0.503
        2.25        0.515        0.494
        2.33        0.469        0.486
        2.42        0.452        0.477
        2.50        0.433        0.470
        2.58        0.414        0.462
        2.67        0.429        0.454
        2.75        0.407        0.447
        2.83        0.423        0.440
        2.92        0.440        0.434
        3.00        0.375        0.427
        3.08        0.391        0.421
        3.17        0.364        0.415
        3.25        0.333        0.409
        3.33        0.350        0.403
        3.42        0.368        0.397
        3.50        0.389        0.392
        3.58        0.412        0.386
        3.67        0.375        0.381
        3.75        0.400        0.376
        3.83        0.357        0.371
        3.92        0.308        0.367
        4.00        0.333        0.362
        4.08        0.364        0.357
        4.17        0.400        0.353
        4.25        0.333        0.349
        4.33        0.375        0.344
        4.42        0.429        0.340
        4.50        0.500        0.336
        4.58        0.400        0.332
        4.67        0.250        0.328


AGAIN WITH SAME FILE? (Y OR N)
N

AGAIN WITH ANOTHER FILE? (Y OR N)
Y
```

(3) BOTH

Under this option, employees in the selected occupation who also were in the selected grade at hire are analyzed.

(4) NONE. RUN WHOLE FILE

This option should be used for files that contain only one occupation. Or, for files that contain only occupations which have been previously analyzed and found to be compatible (see the next chapter).

LPFILE then uses the selected records to calculate a log-probability equation and standard error, both of which are printed out. The user has the option to write out the actual retention pairs and the corresponding curve values.

After each run of LPFILE, the user may recycle either (a) to do another analysis on the same file or (b) to analyze a different file.

Additional information about the development and use of LPFILE can be found in its Technical Analysis and Operation Manual.

CHAPTER 8

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LPTEST

We have said that retention analysis is ideally performed by occupation using longitudinal data. We saw in the preceding chapter cases where longitudinal data techniques are not directly applicable. These cases led to the development of LPFILE. There are also situations where either (a) analysis by individual occupations becomes difficult or (b) analysis within individual occupations is desirable. For such situations, there is a third staffing needs planning computer program known as LPTEST.

When To Use LPTEST

LPTEST deals with the concept of the existence or nonexistence of differences between or within occupations. It statistically compares their retention trends to determine whether or not they may be grouped together for analysis and planning purposes (such as for input to LOGPRO).

There are two types of analysis which can be done using LPTEST. These two areas can be characterized as inter-occupational and intra-occupational.

Inter-occupational analysis: For this type of analysis, LPTEST compares the retention curves of different occupations to see if they have the same or similar retention patterns. Those that do can then be aggregated into a single planning unit.

This analysis capability is useful for organizations consisting of several occupations with only a few employees in each. Such a situation means that there will be only a small number of hires into any one occupation during a year. This in turn means, as we saw in the preceding chapter, that longitudinal analysis for any one occupation would be imprecise. However, it may be that some of these occupations have retention patterns which are similar enough to allow for their grouping together. Such a grouping can be considered as one occupation for longitudinal retention analysis and projection purposes.

It may also be the case that, after having separated out its major occupations for analysis, an organization might want to combine some of its smaller occupations into one or more larger groups which can be analyzed using log-probability techniques.

As a corollary to this process, LPTEST may be used to test the retention patterns of the same occupation at different times. That is, it may compare the occupation's retention curve as calculated from one year's cohort group with that calculated from another year's cohort group. Such a comparison will show whether there has been any change in the occupation's retention pattern over time.

Analysis for a set of grouped occupations will utilize one log-probability equation to represent the entire group.

Intra-occupational analysis: For this type of analysis, LPTEST compares the retention curves of subgroups of the same occupation to see if they have different retention patterns.

A subgroup of an occupation consists of employees from that occupation who fit into any desired category. Some of the categories may be sex, minority status, grade at hire, training received, etc. For example, LPTEST may be used within an occupation to compare the retention patterns of male vs. female employees, minorities vs. non-minorities, GS-5 vs. GS-7 hires, those given special training vs untrained controls, etc. If a difference is discovered, the affected subgroups can be planned for separately (separate log-probability equations will be available).

The technology used to make such comparisons is also useful in other areas of personnel management. There will be more about this in the next chapter.

There is an important point about such comparisons which should be emphasized here. And that is that they should be made for subgroups within the same occupation. It is not valid to say, for example, that there is a difference in the turnover rates of men and women if the rates compared come from different occupations. (Remember that different occupations have different turnover rates.) Such rates should be compared within occupations where both males and females are strongly represented. This will remove the differences in turnover rates which are solely the result of the differences in occupations.

Statistical Differences

LPTEST compares two or more occupational retention curves to determine if all or some of the occupations may be grouped together. When looking at retention trends, LPTEST is searching for statistically significant differences. The underlined phrase leads to two questions:

1. What constitutes a difference?

2. What is meant by "statistically significant"?

Question 1. There are two conditions which alone or in combination can cause occupations to have differing retention trends. They are (a) differences in the first-year turnover rate and/or (b) differences in the log-probability annual loss rate.

Figure 8-1 shows the retention curves for a professional occupation (Curve I) and a clerical occupation (Curve II). It seems obvious just by looking at these two curves that they are different. For one thing, their first year loss rates are decidedly different (17% for Curve I vs. 42% for Curve II). Secondly, Curve II loses 38% (= 58% - 20%) of its cohort from year 1 to year 7, while Curve I loses only 28% (= 83% - 55%). This means that Curve II has a faster log-probability loss rate per year than does Curve I.

Another way of examining the difference between two curves is by plotting them on a log-probability nomograph. This is done for the two sample curves in Figure 8-2. As you can see, Line I starts and remains above Line II. Also, the distance between the two lines is growing larger as time passes. There is a 25% difference between the first-year retention rates of the two lines. This builds to a 36% difference at the seventh year. In other words, the two lines are diverging (getting farther and farther apart).

These kinds of differences can also be seen by looking at the log-probability equations for the occupations tested. For example, the log-probability equations for the two curves in Figure 8-1 (using the first four years of data) are:

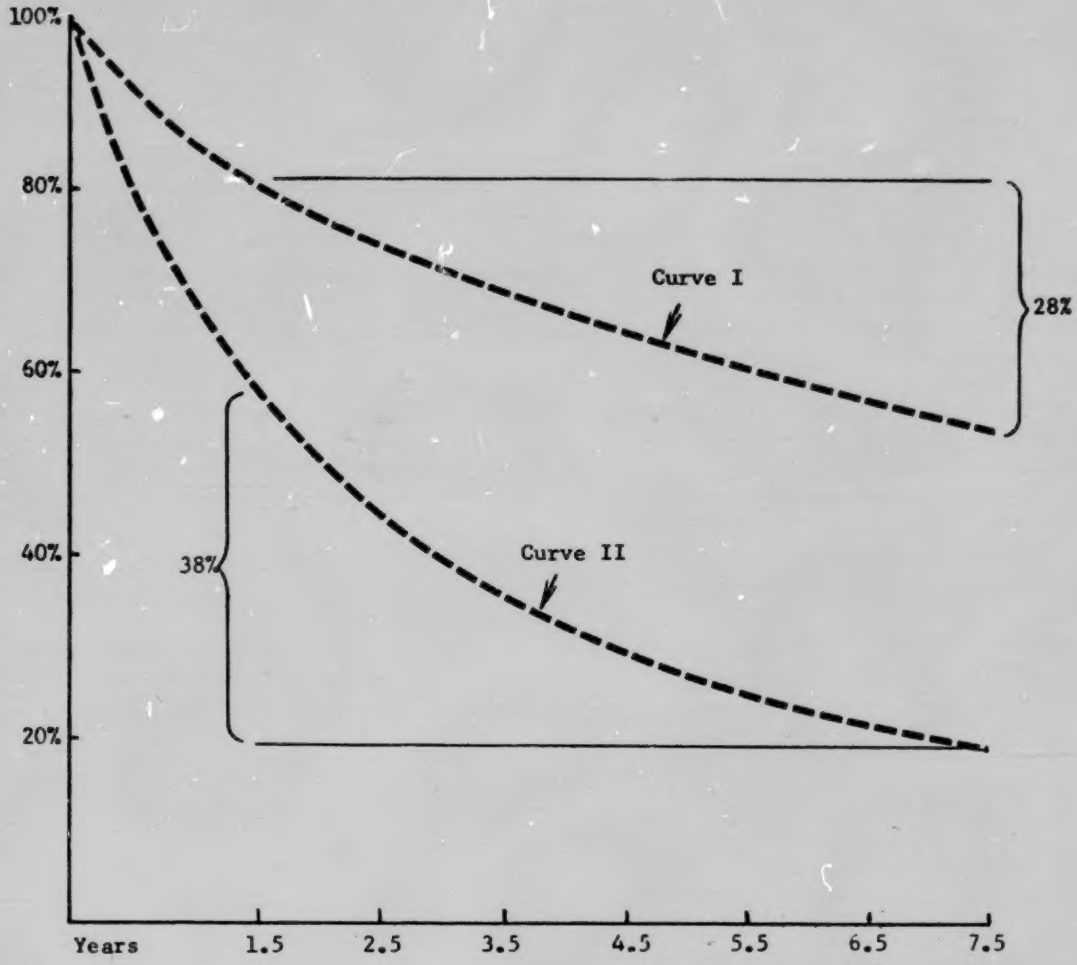
Curve I: $y = 1.15682 - 1.17278x$
Curve II: $y = 0.45692 - 1.49392x$

First, you can see that the A-terms of the equations are quite different. This term by itself represents the normal curve standard deviation value for the time t = 1 (since x = Log t and Log 1 =0). The first A-term, 1.15682, transforms to 88% while the second, 0.45692, translates to 68%. Thus, Curve I has a higher first-year starting point than does Curve II. (This difference is reflected on the nomograph in Figure 8-2. Both of these lines were plotted using a 0.5 averaging factor so that the values on the first vertical line are for t = 1.5.)

The second, or B, term of the equation represents the slope (or steepness) of the log-probability line. Looking again at the lines on the nomograph, you can see that Line II is steeper than Line I. This difference is reflected in the B-terms of the log-probability equations for these two lines, since the absolute value of the B-term for Curve II is larger than that for Curve I.

117

Figure 8-1



Curve I

Years Since Hire	Percent Retained
1.5	82.9
2.5	75.5
3.5	69.8
4.5	65.2
5.5	61.4
6.5	58.1
7.5	55.3

Equation:
Y=1.15682 - 1.17278 X

Curve II

Years Since Hire	Percent Retained
1.5	57.7
2.5	44.5
3.5	36.1
4.5	30.2
5.5	25.8
6.5	22.4
7.5	19.8

Equation:
Y=0.45692 - 1.49392 X

BUREAU OF POLICIES AND STANDARDS, U.S. CIVIL SERVICE COMMISSION

LOG-PROBABILITY NOMOGRAPH FOR PROJECTING WORKGROUP RETENTION (AND LOSS)

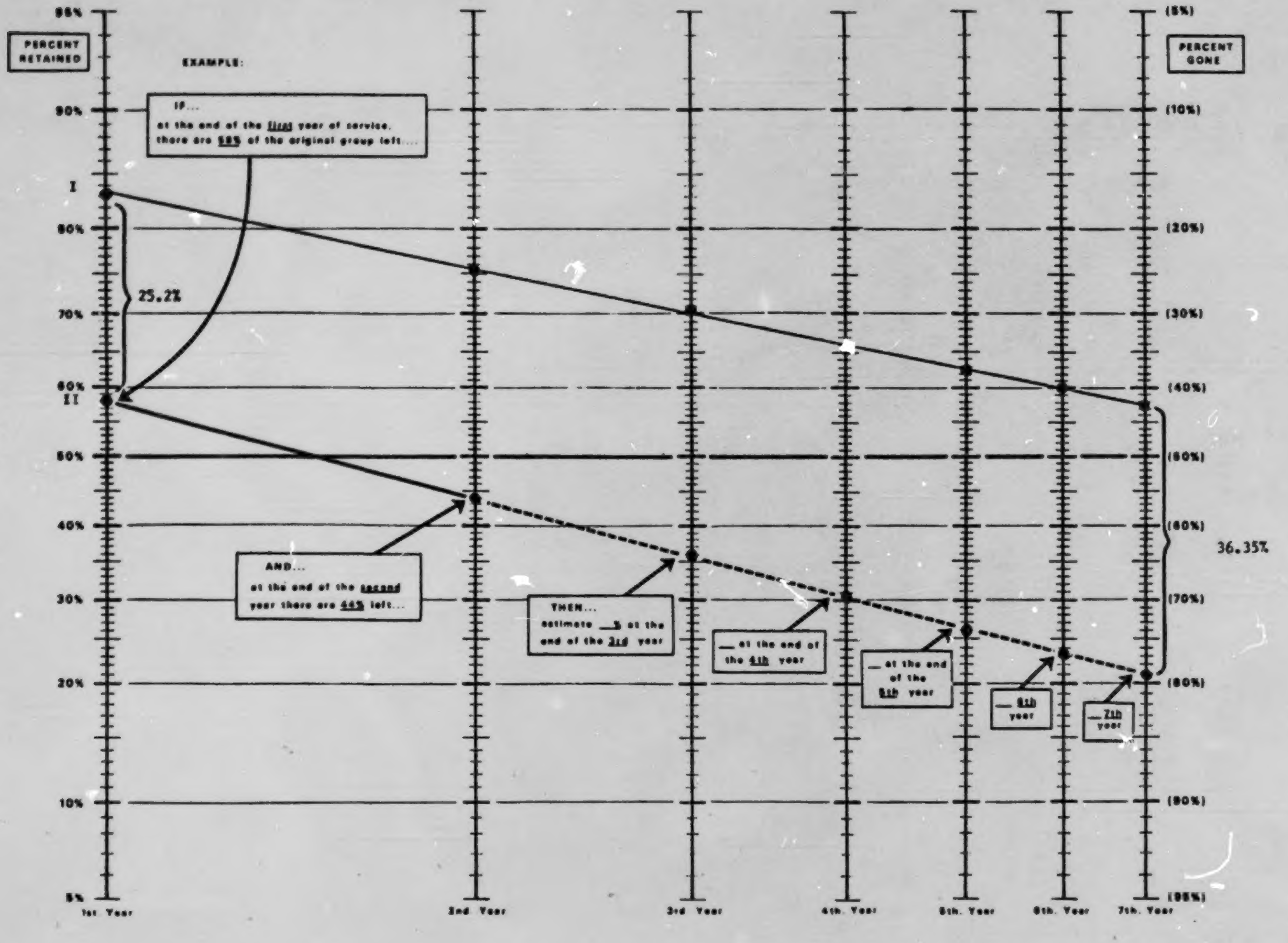


Figure 8-2

Such inspection techniques are useful in determining occupational differences when occupations have obviously different retention patterns. However, most occupations that will need to be compared will be much closer in trend than the two sample occupations. The statistical technique used by LPTEST will be useful in these cases.

In addition, LPTEST may be used to test more than two occupations, which would be difficult to accomplish visually.

Question 2. Just the existence of a difference, however, is not enough evidence to say that the tested occupations may not be grouped together. It is necessary to go one step further and test whether the difference is statistically significant.

First it should be said that when we compare occupations using LPTEST, we are testing to see if the individual sets of retention data form samples which come from the same overall population. If they do, then they may be grouped.

No two sets of data are ever going to be exactly alike. There will always be some difference which is due to the action of chance factors. What is needed is a criterion to determine when the occurrence of a difference means that the compared sets come from different populations. This criterion is known as statistical significance. Statisticians have defined differences as being statistically significant when the probability that their occurrence could be attributed to chance is 1 in 20 (or less).

To determine what probability value is associated with a given difference, the difference must be quantified. This is done by first setting up an hypothesis which assumes that no difference exists (the "null" hypothesis) and then trying to disprove it using an appropriate statistical test. This test will result in a number which will be associated with a predetermined unique probability value.

By definition, if this unique pre-determined probability value is less than 0.05 (= 1/20), then a statistically significant difference will exist. If the probability value is less than 0.01, then the difference is considered to be highly significant. These probability values are also known as significance levels.

Statisticians have worked out formulas and developed tables which enable one to determine what probability value is associated with the result achieved by using a given statistical test. These tables can be found in any statistics book.

The Statistical Test. The selection of a statistical test should be carefully made. There are many such tests and each one should be used only in the situation(s) for which it was designed.

For its comparisons, LPTEST uses the "F-statistic-test." This test compares two variances (by dividing the larger variance by the smaller) to get a value which is known as "F." If this value is large enough, then the difference between the two variances is considered to be significant.

The significance level associated with a given value of F can be inferred from statistical tables or approximated using mathematical equations. LPTEST contains a subroutine using one such mathematical approximation. This subroutine uses a calculated value for F plus what are known as the associated degrees of freedom figures for each of the two previously-calculated variances. Generally speaking, degrees of freedom is defined as the number of data items in a sample minus one. In LPTEST, the number of data items equals the number of items from which a variance is calculated.

The LPTEST Method

LPTEST makes use of longitudinal data to compare occupational retention trends. These data are in the same format as those which are entered into LOGPRO. Thus, you need to enter, for each occupation to be tested, the number in the starting group and the number or percent retained at later points in time. LPTEST also uses the same statistical transformations and basic methodological assumptions that are used by LOGPRO.

LPTEST can also use data which come from an LPFILE program run. This is done by using LPFILE to compute an equation for the occupations concerned and replying "YES" to the option to write out the actual and curve values. Then you can select from these lists actual percentage values for a few time values (using the same time values for each occupation). These can then be entered into LPTEST. Since the retention (or y-) values will already be percentages, any convenient values can be entered into LPTEST as starting group figures.

As was discussed at the beginning of this chapter, LPTEST type data are either grouped by occupation (for inter-occupational analysis) or by subgroup within an occupation (for intra-occupational analysis). If any set of groups or subgroups is found to be compatible, they may then be combined for further analysis. For example, combined longitudinal data can be entered into LOGPRO for analysis and projection.

Any group of occupations which is found to be incompatible can be "regrouped"--i.e., one or more of the occupations can be removed from the test group and the rest can be run through the program again. This process can be repeated as often as the user wishes.

To test the inputted longitudinal data points for any differences, LPTEST sets up the null hypothesis that there are no differences among the retention trends of the inputted occupations. Then it calculates a value for the F-statistic and determines the probability associated with that value of F.

To determine a value for the F-statistic, LPTEST uses what is known as an analysis of variance technique. This involves using all of the inputted longitudinal data to compute two specific variance figures:

Va = The variance among the inputted groups

Vw = The variance within the inputted groups

The first variance, Va, represents the differences among the inputted groups (or how the groups differ from each other). The second, Vw, estimates the variation that occurs within each group. The F-statistic is calculated using the formula:

$$F = \frac{Va}{Vw}$$

If the null hypothesis is correct, then there should be little or no difference between the two calculated variance figures. In other words, the value of F would be fairly close to 1. However, if a significant difference exists, then the variation among the inputted groups will be considerably larger than the variations within the individual groups. If this is so, then the value of F will be large.

The degrees of freedom figure associated with Va (Na) is equal to the number of groups tested minus one. The degrees of freedom figure associated with Vw (Nw) is the sum of all the longitudinal data points entered minus one. For example, if three occupations are tested and each occupation has four years of retention data, then

$$Na = 3 - 1 = 2$$

$$Nw = (3 \times 4) - 1 = 11$$

The values of F, Na, and Nw are entered into the F-approximation subroutine (FTEST) which returns the associated probability value to LPTEST. If this value is less than 0.05, then the differences among the inputted groups is statistically significant.

LPTEST also calculates the log-probability equation for each inputted group. If it is determined that one or all of the inputted groups should be analyzed separately, then the log-probability equation(s) calculated by LPTEST can be used in any further analysis. If, on the other hand, the tested groups are found to be compatible, LPTEST calculates the log-probability equation which represents the combined groups.

Sample Run of LPTEST

LPTEST is designed to accept longitudinal retention data for up to twenty occupations or groups. These data are entered during the first- or information-phase of an LPTEST run. A user may select all or any subset of the entered groups for testing. The test results are printed out during the analysis output phase of LPTEST.

During the information input phase of a run, LPTEST asks for the following pieces of information:

(1) The number of groups (e.g., occupations) which are to be tested.

(2) The occupation or series code for each inputted group. (If subgroups of one occupation are being entered, each one will need a numeric code.)

(3) The number of time (or x-) values for which there are retention data. Each group tested must have the same number of x-values.

(4) The value of each x-value. Each group tested must use the same x-values.

(5) The starting population for each cohort.

(6) Whether the retention (or y-) values are in the form "number of employees retained" or "percent of employees retained."

(7) The actual retention values.

The data for item (7) are entered by x-value. That is, all of the y-values associated with a given x-value are entered on one line and separated by commas. The order and number of the retention values on a line is the same as the order in which the occupation codes were entered.

In addition, LPTEST asks for information to determine which of the inputted groups are to be tested at this time. First, it asks if the user wishes to test (1) all of the groups or (2) only some of the groups. If all of the groups are to be tested, LPTEST goes directly to its analysis sequence. If only some of the groups are to be tested, LPTEST queries the user as to how many groups are to be tested and which ones they are.

Figures 8-3 through 8-7 show a sample analysis sequence using LPTEST. Five different occupations are entered and comparisons are made using three different subsets of these occupations. The sample data used are:

OCCN CODE	N	PERCENT RETAINED AFTER 1 Yr.	2 Yrs.	3 Yrs.
322	30	58.00%	44.00%	36.00%
312	25	52.73	40.00	32.73
201	62	83.20	74.50	70.30
212	50	80.00	70.00	61.00
1520	15	92.00	87.00	85.00

Figure 8-4 shows the order in which percent retained values are entered: first, each occupation's "1 Yr." values are entered (in the same order as the occupation codes were entered), then, all of the "2 Yrs." values, etc.

Since the sample occupations are a diverse mixture, they would seem to fall into at least two obvious groups: clerical (322 and 312) and PAT[1] (201, 212, and 1520). Using this breakdown, the first test made was a comparison of 322 and 312. (Note in Figure 8-5 that to make this selection, we entered "1, 2." These are the values that were assigned to these groups during the entry of occupation codes.)

The result of this first test can be found in Figure 8-5. As you can see, this output shows which occupations have been tested and states that they may be grouped together. The log-probability equation for each group tested is also printed out. In addition, since these two occupations were found to be compatible, the log-probability equation of their combined retention values is printed (under the heading "Total Group Equation").

Next, a test was made to compare 201, 212, and 1520. The analysis output in Figure 8-6 gives the result of this test: these three occupations may not be grouped. Since it seems likely that occupation 1520 is the one which is gumming up the works, this occupation was removed from the group and another test was made. The result of this last test is shown in Figure 8-7. The two occupations, 201 and 212, may be grouped so their total group equation is printed out.

So, from this analysis of five occupations, we came up with three groupings: 322 and 312; 201 and 212; 1520. These groupings may be used in further analyses involving these occupations or the occupations may be analyzed individually. Either way, the log-probability equations needed are produced by LPTEST.

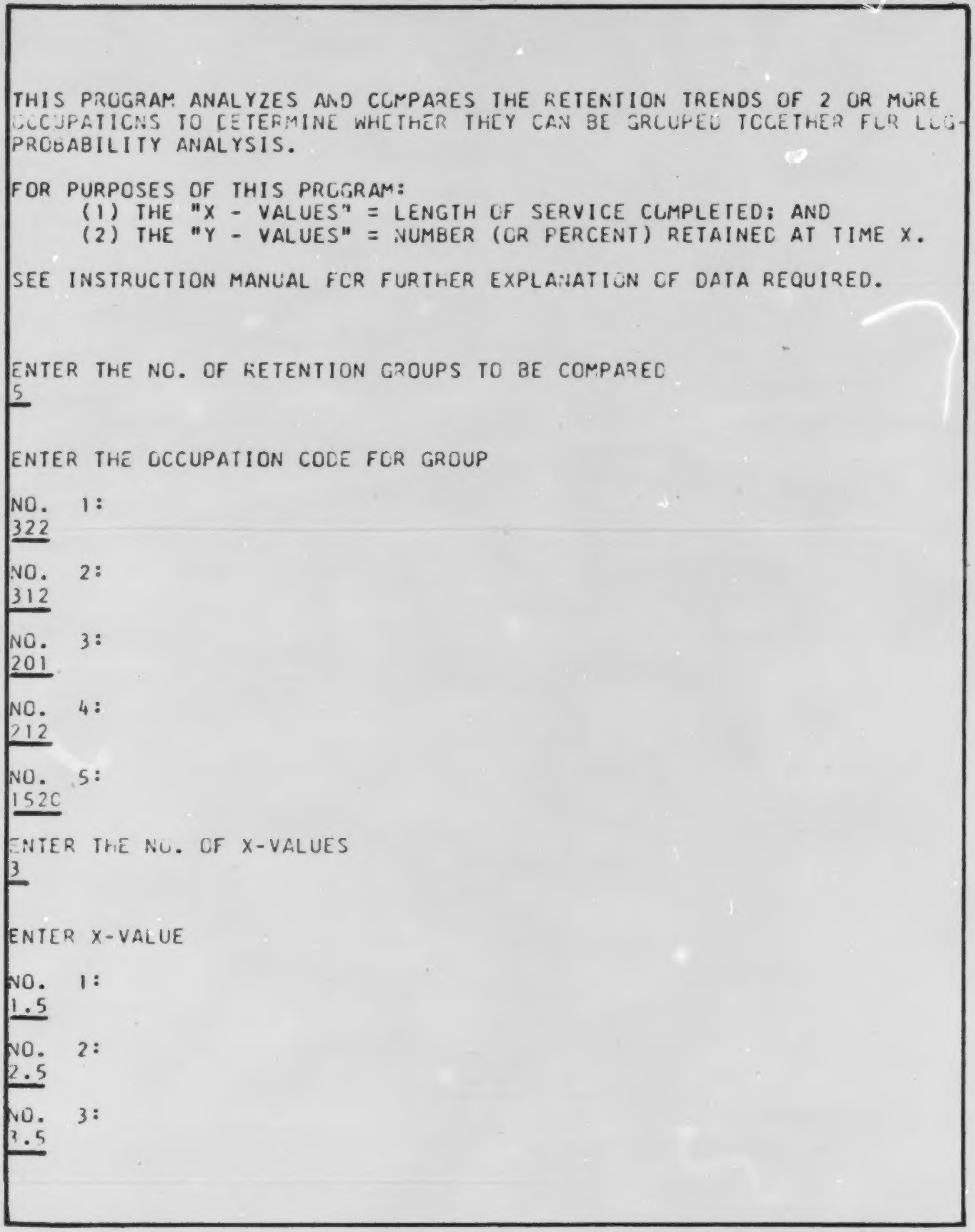
After each run of LPTEST, the user may recycle back to the beginning to perform another comparison by either using the same occupations or inputting new occupations.

Additional information about the development and use of LPTEST can be found in its Technical Analysis and Operation Manual.

---

[1] PAT is shorthand for Professional, Administrative and Technical occupations.

Figure 8-3

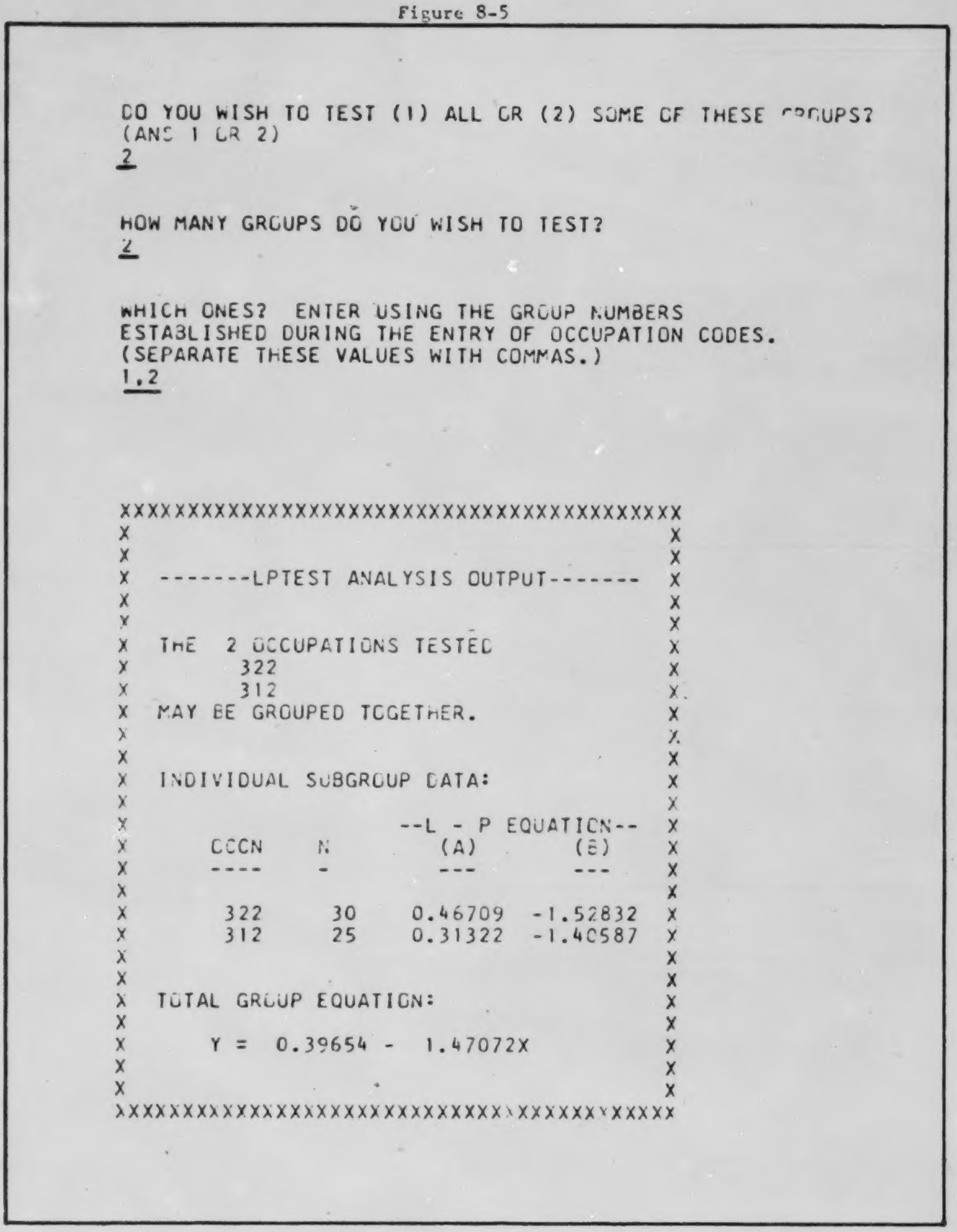


```
THIS PROGRAM ANALYZES AND COMPARES THE RETENTION TRENDS OF 2 OR MORE
OCCUPATIONS TO DETERMINE WHETHER THEY CAN BE GROUPED TOGETHER FOR LOG-
PROBABILITY ANALYSIS.

FOR PURPOSES OF THIS PROGRAM:
      (1) THE "X - VALUES" = LENGTH OF SERVICE COMPLETED; AND
      (2) THE "Y - VALUES" = NUMBER (OR PERCENT) RETAINED AT TIME X.

SEE INSTRUCTION MANUAL FOR FURTHER EXPLANATION OF DATA REQUIRED.


ENTER THE NO. OF RETENTION GROUPS TO BE COMPARED
5

ENTER THE OCCUPATION CODE FOR GROUP

NO.  1:
322

NO.  2:
312

NO.  3:
201

NO.  4:
212

NO.  5:
1520

ENTER THE NO. OF X-VALUES
3

ENTER X-VALUE

NO.  1:
1.5

NO.  2:
2.5

NO.  3:
3.5
```

Figure 8-4

```
ENTER STARTING POPULATICN (N) FOR GROUP

NO.  1:  ( 322)
30

NO.  2:  ( 312)
25

NO.  3:  ( 201)
62

NO.  4:  ( 212)
50

NO.  5:  (1520)
15

ARE Y-VALUES IN
   (1) NUMBER OR
   (2) PERCENT FCRM?
   (ANS 1 OR 2)
2

IN DECIMAL FORM, SEPARATED BY COMMAS
ENTER THE Y-VALUES CORRESPONDING TO X =

 1.5000:
.58,.5273,.832,.80,.92

 2.5000:
.44,.40,.745,.70,.87

 3.5000:
.36,.3273,.703,.61,.85
```

Figure 8-5



```
DO YOU WISH TO TEST (1) ALL OR (2) SOME OF THESE GROUPS?
(ANS 1 OR 2)
2

HOW MANY GROUPS DO YOU WISH TO TEST?
2

WHICH ONES?  ENTER USING THE GROUP NUMBERS
ESTABLISHED DURING THE ENTRY OF OCCUPATION CODES.
(SEPARATE THESE VALUES WITH COMMAS.)
1,2



XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
X                                      X
X                                      X
X  -------LPTEST ANALYSIS OUTPUT-------  X
X                                      X
X                                      X
X  THE  2 OCCUPATIONS TESTED           X
X        322                           X
X        312                           X
X  MAY BE GROUPED TOGETHER.            X
X                                      X
X                                      X
X  INDIVIDUAL SUBGROUP DATA:           X
X                                      X
X                     --L - P EQUATION--  X
X      OCCN     N        (A)        (B)   X
X      ----     -        ---        ---   X
X                                      X
X       322    30    0.46709   -1.52832   X
X       312    25    0.31322   -1.40587   X
X                                      X
X                                      X
X  TOTAL GROUP EQUATION:               X
X                                      X
X      Y =  0.39654 -  1.47072X        X
X                                      X
X                                      X
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Figure 8-6

```
DO YOU WISH TO TEST ANOTHER SUBSET OF THESE GROUPS (Y OR N)?
Y


DO YOU WISH TO TEST (1) ALL OR (2) SOME OF THESE GROUPS?
(ANS 1 OR 2)
2


HOW MANY GROUPS DO YOU WISH TO TEST?
3


WHICH ONES?  ENTER USING THE GROUP NUMBERS
ESTABLISHED DURING THE ENTRY OF OCCUPATION CODES.
(SEPARATE THESE VALUES WITH COMMAS.)
3,4,5




XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
X                                       X
X                                       X
X  -------LPTEST ANALYSIS OUTPUT-------  X
X                                       X
X                                       X
X  THE  3 OCCUPATIONS TESTED            X
X        201                            X
X        212                            X
X       1520                            X
X  MAY NOT BE GROUPED TOGETHER.         X
X                                       X
X                                       X
X  INDIVIDUAL SUBGROUP DATA:            X
X                                       X
X                    --L - P EQUATION--  X
X      OCCN    N        (A)        (B)  X
X      ----    -        ---        ---  X
X                                       X
X       201    62   1.15866   -1.18236  X
X       212    50   1.11498   -1.52008  X
X      1520    15   1.57040   -1.02254  X
X                                       X
X                                       X
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Figure 8-7

```
DO YOU WISH TO TEST ANOTHER SUBSET OF THESE GROUPS (Y OR N)?
Y


DO YOU WISH TO TEST (1) ALL OR (2) SOME OF THESE GROUPS?
(ANS 1 OR 2)
2


HOW MANY GROUPS DO YOU WISH TO TEST?
2


WHICH ONES?  ENTER USING THE GROUP NUMBERS
ESTABLISHED DURING THE ENTRY OF OCCUPATION CODES.
(SEPARATE THESE VALUES WITH COMMAS.)
3,4




XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
X                                      X
X                                      X
X  -------LPTEST ANALYSIS OUTPUT-------  X
X                                      X
X                                      X
X  THE  2 OCCUPATIONS TESTED           X
X        201                           X
X        212                           X
X  MAY BE GROUPED TOGETHER.            X
X                                      X
X                                      X
X  INDIVIDUAL SUBGROUP DATA:           X
X                                      X
X                  --L - P EQUATION--  X
X      OCCN    N        (A)       (B)  X
X      ----    -        ---       ---  X
X                                      X
X       201   62    1.15866  -1.18236  X
X       212   50    1.11498  -1.52008  X
X                                      X
X                                      X
X  TOTAL GROUP EQUATION:               X
X                                      X
X      Y =  1.13857 -  1.33711X        X
X                                      X
X                                      X
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

CHAPTER 9

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

GS810

The Need for Analysis of Advancement

Having covered the detailed methods of analyzing and projecting turnover, the logical next step is to show how to utilize this technology in estimating hiring needs. At this point, however, while the detailed discussion of the LPTEST program is still fresh, we want to show how an adaptation of LPTEST techniques for determining differences can be applied to the analysis of grade-advancement patterns.

Advancement, of course, is one of the most fundamental of all personnel movements. And the goal of advancement based on merit and fitness for the work of the service is one of the most fundamental goals of the Merit System itself. Apart from its intrinsic interest, however, there are a number of specific purposes for developing methods of analyzing advancement patterns.

One purpose is for occupational studies. An occupation's advancement pattern is one of that occupation's most characteristic and distinctive features. Being able to analyze advancement patterns and to distinguish between the advancement patterns of different occupations are important aspects of occupational analysis work.

A second purpose of analyzing occupational advancement patterns is for use in employee career counseling activities and--along with data on inter-occupation mobility trends--in the establishment of lines of promotion for setting up merit promotion plans.

A third reason is for the purposes of research. Being able to analyze and compare advancement patterns for different groupings would give us a powerful tool, for example, for comparing the effects on advancement of such factors as differences in personnel characteristics, or the effects of different kinds of training programs. Advancement of employees who have advanced degrees, for example, versus those who do not. Or advancement of employees given special post-entry training versus those who were not. And so on.

A fourth purpose for analyzing and comparing advancement patterns is to identify individuals and/or groups whose advancement is significantly above or below the norm for their occupational group. Identifying employees whose advancement is much above average, for example, can be very useful in programs for identifying potential future executives. Identifying employees with below-normal advancement, on the other hand, can be useful in remedial training, performance evaluation, and similar programs.

Finally, GS810 can be of use in equal employment opportunity programs, in comparing advancement patterns of minority groups with patterns of non-minorities.

Equal employment opportunity, of course, is an important responsibility both of government--Federal, State and local--and of private industry as well. Most major employers, in fact, devote substantial and continuing efforts toward the establishment and improvement of their EEO programs.

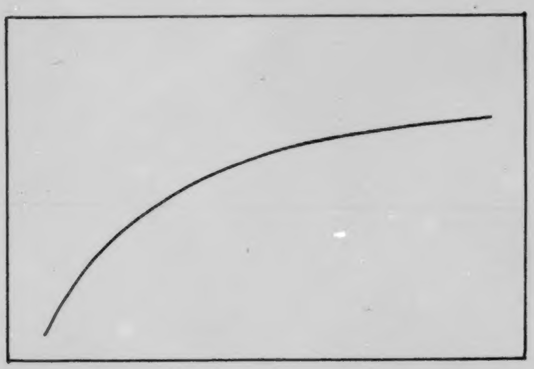
Data analysis usually is a key element of all of the above-mentioned programs. It provides objective means of assessing program progress. It can help to pinpoint program areas which may be in need of improvement. It can show the effects of past actions. And it can indicate the likely future effects of current actions.

We believe that in LPTEST we have an analytical technology which can be of significant help in such analyses.

Barriers To Analysis

There are truly formidable barriers, however, to applying LPTEST techniques to the analysis of advancement data. Let us cite just a few.

First, we have established through research which has followed group grade advancement over many years that the mean grade of a given group of, say, GS-5 hires, rises over time along a log-normal curve. This is just like the L-P (Log-Probability) turnover curve described earlier except that it goes up instead of down, as in the following illustration.



The significance of the fact that this is a rising curve lies in the necessity--which was explained earlier in the transformation of L-P curve data--for converting y-axis data into "Percent of Maximum Possible Value of y" and thence into standard deviations. For L-P curves, we could do this because we knew the maximum possible value of y: the number of employees we started out with. Because advancement curves are rising curves, however, we have no way to determine what the maximum possible value of y might be. Thus, we have no way of converting our data into the "standard deviation" form needed for least-squares fitting. The curve form used in LPTEST, therefore, cannot be used for advancement data.

Those with economics backgrounds might at this point suggest that we substitute for the L-P curve form one of the standard growth curves used in economics, such as the Pearl-Reed logistic or the Gompertz:

(1) $1/y = a + bc^x$ (logistic); or
(2) $\log y = \log a + \log b(c^x)$ (Gompertz)

These curves are such, however, that they can be fitted to data only by approximate techniques; least-squares fitting is impossible. If they were used, we could then not use the F-test of variance to permit comparison of one curve relative to another.

The second major barrier to the use of the LPTEST technique is that of the availability of data. The L-P curve technique would require longitudinal data on workforce advancement trends for, say, ten to twenty years past. Such data are simply not available. Thus, here too, the techniques of LPTEST cannot be directly applied.

Finally, any analysis of advancement trends must have some means of scaling jobs by grade level; i.e., into numbered intervals, rather than the continuous variable y-values used in LPTEST. GS810 uses the General Schedule grading system used in Federal white-collar employment. The GS system is not used, however, in State or local government or in private industry. To be useable outside the Federal government, then, some other means of scaling job levels--y-values--must also be provided for.

GS810 Features

Formidable as these three barriers are, they are not insurmountable. To overcome them, four major adaptations of LPTEST techniques have been developed. These give our advancement program GS810 four characteristic features which distinguish it rather sharply from LPTEST and which should be thoroughly understood. These are as follows.

First, GS810 uses census-type data, rather than longitudinal data, in its calculations. That is, all of the data used in GS810 are collected at one point in time, rather than over several successive time points.

133

We can illustrate the difference this way. Suppose you wanted to study children's growth patterns during their grade school years. If you were to use the longitudinal method, you would take a particular group of children just entering grade school, measure their heights at that point, and re-measure them annually thereafter until they left grade school.

To use the census method, on the other hand, you would measure all the children in the school at the same time. Then you would group data from all first-graders together, all second-graders, all third-graders and so on, and average the heights in each grade separately. The resultant curve formed by plotting these averages would look much like the curve that would be gotten from longitudinal data, except that each year's observations would be of different children.

The second major difference between LPTEST and GS810 is that GS810 uses a different curve form for fitting to the data. The form used is a variety of exponential:

$$y = a\,x^{b}$$

The reasons that this form is used can be summarized quite briefly. First, by means of logarithms, it can be converted into the simple linear form:

$$\log y = \log a + b \log x$$[1]

This form can be fitted to data by straightforward least-squares techniques, and the F-test procedures similar to those used in LPTEST can be used in GS810 with full validity.

Second, when the total time span covered by the data is held to not more than about 10 years--as compared to the 30-40 years that might be theoretically possible in an old-time organization--this curve form fits grade advancement data so closely that it is statistically virtually indistinguishable from what could be gotten using the log-normal curve form.

The third major difference between LPTEST and GS810 is that in GS810 the y-values data are in grouped-data form, rather than in continuous variable form. In L-P curve fitting, you will recall, the value of y--the fraction of the starting group still present--was a continuous variable which could take any value between 0 and 1.0 (0% to 100%). Thus, we could--and did--get values like 0.1275 (12.75%), 0.6733 (67.33%), and so on.

1/ This form of linear equation is referred to as "log-log" since both the x- and y-values are transformed to their corresponding logarithms.

134

Grade level figures, however, are not continuous variables. There are no fractional grades. There are only a limited number of grades into which the full spectrum of job difficulty must be grouped. Such groupings obviously ignore fine differences in job difficulty. Thus, a job that barely reaches the GS-3 level of difficulty is "GS-3," not "GS-3.01." By the same token, another job, which falls just a hair short of reaching the GS-4 level, is also "GS-3," not "GS-3.99."

Each grade-level number, in other words, stands for an interval. "GS-3" stands for "GS-3.00 to GS-3.99." "GS-4" stands for "GS-4.00 to GS-4.99." And so on. Grades, then, are just like the ranges that are used when grouping continuous variable data together according to intervals.

To return to our grade-school example, for a moment, suppose we took all second-graders whose height we measured and we made up a grouped data table showing how many there are in the 46-47.99" interval (3 feet 10 inches, but not 4 feet), how many are in the 48-49.99" interval, and so on. To make up a "grouped-data average," of course, we would take the number of children in each interval, multiply this times the midpoint of each interval (47.00", 49.00", etc.), and then divide the overall sum of these products by the total number of children to get our desired average.

GS810 handles grades the same way. Since "GS-3" includes all jobs whose difficulty falls in the range "3.00 - 3.99," the midpoint of the "GS-3" interval is weighted "3.50." Similarly, the "GS-4" range midpoint is "4.50." And so on.

The "length-of-service" dimension is treated in exactly the same fashion. The employees who are in their first year of service fall in the "0.0 - 0.99" interval. The midpoint of this interval is "0.50." The midpoint of the "1.00 - 1.99" years of service interval is "1.50." And so on.

In GS810, then, all computations are done from grouped-data tables, with both x-axis (length of service) values and y-axis (grade) values made up of intervals. Such an 8 row by 10 column tabular format is an "8 by 10 matrix," in computer parlance. (See sample below.) This is where the "810" part of GS810's name comes from.

The "GS" part of the name comes from the fact that GS810 automatically provides correct weights for the two most common grade progression patterns of the Federal "General Schedule" (GS) occupations: GS-1/8 and GS-5/15. (The program also permits entry of other weight patterns, if desired, for non-GS or non-Federal occupations.)

The fourth major difference is that although LPTEST and GS810 both use the F-statistic to determine if a difference exists, the two programs use different methods to obtain the variances which are compared.

In LPTEST, the variances which are tested are those within and between columns. In GS810, on the other hand, the data for each individual subgroup are added together and a log-log least-squares line is fitted to the summed (or total group). The resulting equation represents the overall group norm. In addition, log-log least-squares lines are fitted to each individual subgroup and three types of variances are calculated:

1. The variance of the total group data from the total group equation;
2. The variance of each subgroup's data from the total group equation; and
3. The variance of each subgroup's data from its own equation.

All of these variances are calculated utilizing differences between actual and curve values as described in Chapter 3.

Three kinds of comparison tests are made using different combinations of these variances. A value for the F-statistic is calculated for every test. The tests are discussed later.

Preparation of Data

To assemble the data needed for a GS810 analysis, then, we would proceed like this. First, we would take all employees who have ten years or less of service and for each such employee, we would make up a card showing:

(1) occupation
(2) sex
(3) minority status [2]
(4) length-of-service since entry on duty
(5) grade

Then, we would sort these cards into stacks by occupation--one stack per occupation. The number of stacks we get will determine the number of GS810 occupational analyses we can do.

To prepare an occupation for a GS810 analysis, we then sort each occupational stack into four piles:

Minority Male (MM)
Minority Female (MF)
Non-minority Male (NM)
Non-minority Female (NF)

---

[2] For other phases of advancement research, data elements (2) and (3) may be replaced or supplemented by other test characteristics (e.g., receipt of post-entry training, age-at-hire, educational level, etc.)

We then take each pile and sort it by length-of-service into ten groups: 0-0.99 years, 1.00-1.99 years, etc., and sort each of these in turn by grade. Then, we count the number of cards in each block of our 8 x 10 table and we record the results in matrix form like this (entering zeroes in each block where no employees were found):

0	0	0	0	0	0	0	0	0	0
0	0	0	1	2	7	7	1	3	3
0	0	0	1	4	4	8	2	4	7
1	3	9	13	24	53	21	6	5	3
31	42	51	40	48	48	28	4	6	6
119	128	109	19	27	23	17	2	2	4
82	30	10	2	1	5	4	1	3	0
10	3	0	0	0	1	0	0	0	0

Following the GS810 technical directions, we enter these data into the computer as named files. For occupation 999, for example, we would have files named "NM999," "MF999," and so on.

From this point on, the complete analysis job will be done entirely by GS810 in the pattern desired by the program operator.

Using GS810

Since the GS810 Operation Manual provides a detailed, step-by-step description of GS810's features and options, we will not repeat these details here. Rather, we would like to discuss the principal things that GS810 does and how it can help the analyst identify the nature and source of the kind of problems that are typically encountered.

At the outset, it must be clearly understood that GS810 must be used only for the kind of comparisons for which it was designed: comparisons of the overall norm of the occupation of which the subgroups are a part. To put it another way, GS810 cannot be used to compare any subgroup with the norm of any group of which the subgroup is not a part.

The reason for this is inherent in GS810's methodology. It is an absolute requirement for GS810's use of the F-test that all of the data for the subgroup being tested must have been used in computing the overall norm curve and variance against which the subgroup is tested. Any departure from this requirement, however slight it may appear, destroys the validity of comparison results.

With this preliminary caution firmly in mind, let us now look at what GS810 does in its comparisons and how it can help the analyst pinpoint problem areas.

To begin with, GS810 tests each subgroup in three ways. Once to test the subgroup's grade-advancement trend against the occupational norm. Once to compare the width of the subgroup's skill spread above and below the advancement trend line with the skill spread of the occupation as a whole. And once to compare the total grade-time pattern of the subgroup with the total grade-time pattern of the overall occupation.

More specifically, the first test is an evaluation of the value of the ratio:

$$F_1 = \sigma^2_{SG} \Big/ \sigma^2_{SS}$$

where

σ^2_{SG} = variance of subgroup from overall group advancement curve

σ^2_{SS} = variance of subgroup from curve computed from subgroup data only

This test is designed to answer the specific question, "Is the grade advancement curve of this workforce subgroup significantly different from the overall group's norm?" If the answer to this question is "Yes," the program then determines whether the subgroup curve is:

"Higher" -- Both A and B terms of the subgroup curve equation are above the corresponding group norm equation values;

"Lower" -- Both A and B terms are below the group norm values; or

"Different" -- One value is above, one below.

When a "significant difference" is found on this test, the analyst should carefully check the equation values in the "Results" table and/or plot out on graph paper both the subgroup curve and the group norm curve. If the "A" term values are substantially equal, and the "B" term values are conspicuously unequal, then a plot of curve values will show the curves starting at or near a common point and then diverging progressively over time. This is the classic pattern of a group which is receiving clearly differential treatment as compared to the norm.

If, however, the "A" term difference is as great or greater than the "B" term difference, the case is an ambiguous one. It could be the result of differences in the grade distribution of new hires. Or, it could be the result of marked changes in hiring patterns during the ten-year period of the sample. Or, it could be the result of a number of other factors. When the two respective curves, then, converge or intersect, rather than diverge from a common starting point, further investigation is needed to determine the causative factors. No clear-cut, direct finding is possible.

The second test in the GS810 technical sequence is an evaluation of the ratio:

$$F_2 = \sigma^2_{GG} \Big/ \sigma^2_{SS} \quad \text{OR} \quad \sigma^2_{SS} \Big/ \sigma^2_{GG}$$

(whichever is greater), where

σ^2_{GG} = variance of overall group from group equation

σ^2_{SS} = variance of subgroup from subgroup equation

In this test, we remove advancement from both subgroups and overall group data to test the question, "Is the spread of subgroup employee grade (skill) levels above and below the grade advancement curve significantly different from the overall group norm?" If the answer to this question is "Yes," the program then determines whether the subgroup spread is "greater" or "smaller" than that of the overall group.

Generally speaking, a finding that the subgroup grade spread is "smaller" is more common than the finding of "greater," particularly when the subgroup involved represents a minority. In most cases, the "smaller" spread means that subgroup employees are closely clustered around the grade advancement curve with few or no employees being much above or below it, even in the first few years of service. Generally, this indicates that relatively few applicants who are qualified for above-basic-entry grades are available for hire in the labor market. Thus, a greater proportion of minority hiring is at the basic entry grades than is true of non-minority groups.

Where the finding is "smaller," then, additional study of the proportion of subgroup above-entry eligibles in the labor market is needed to establish the nature of the problem.

A finding that the subgroup's grade spread is "greater," on the other hand, almost always stems from the subgroup's getting a greater than normal proportion of its entrants by in-service accessions--e.g., the promotion of long-service clerical employees into professional entrance grades. This can be recognized in the subgroup's grade-time matrix by a greater than expected proportion of employees showing up in the entry grade(s) in the later years of service (6th, 7th, etc.).

A finding of "greater" spread, that is, is a common result of effective upward mobility programs. This effect can be readily cross-checked by redefining the grade-time matrix substituting "length of service in this occupation" for "length of service since hire." If the "greater" finding is the result of in-service hiring factors, this will cause the difference to disappear.

The third and final test of GS810 is a test of the overall subgroup matrix against the overall group matrix. This involves the valuation of the ratio:

$$F_3 = \sigma^2_{SG} \Big/ \sigma^2_{GG}$$

where

σ^2_{SG} = variance of the subgroup from the overall group equation

σ^2_{GG} = variance of the overall group from the overall group equation

This test is designed to answer the overall question, "Does the overall subgroup grade/time pattern which results from the combined effects of (a) the subgroup grade-advancement curve and (b) the subgroup's grade spread, significantly differ from the overall pattern of the occupation as a whole?"

The findings possible on this question are either "different" or "not different." In almost every case of a "different" finding, a "significant difference" finding will have shown up on one or both of the first two tests. Essentially, then, the third test measures the overall effect on the subgroup grade distribution pattern of the curve and variance factors which were tested separately in the first two tests.

Putting it another way, if the third test results in a "significant difference" finding, then the subgroup can be expected to show differences by other tests as well--average grade of subgroup vs average grade of occupation, percentage of subgroup population reaching upper grade levels as compared to overall occupation population, and so on.

Further, if the subgroup distribution shows up as significantly different for employees with up to ten years of service, it is highly likely that these differences will continue to be present for the foreseeable future.

What a "significant difference" finding on this third test does not automatically mean, however,--and this must be emphasized in the strongest terms--is that these differences must have resulted solely and entirely from discriminatory treatment. On the contrary, as we saw earlier, differences can result from effective upward mobility programs, unusually high levels of occupational advancement, difficult labor market conditions, and so on. Differences can result from discriminatory treatment, of course. But the finding of a "significant difference" on the third test by no means equals a finding of discrimination.

Another caution is also in order, this one more general. If the population under study includes employees who were hired long before EEO programs began to have a real impact on recruitment and hiring practices, past discrimination could be misinterpreted as an indicator of present discrimination in the occupational series studied, thus overestimating the extent of present discrimination. One way to avoid this is to provide for separate analysis of persons hired more recently--say, since 1969 (E. O. 11478) or 1972 (the EEO Act).

Conclusion

Let us conclude this chapter by pointing out once again that GS810 is a powerful analytical tool for personnel specialists and other interested officials to use in evaluating and comparing occupational and subgroup advancement rates. It will do many things for you and do them accurately and objectively.[3] But it does not and cannot substitute for the reasoned judgment of the analyst. It can be an invaluable guide for pinpointing areas of inquiry. But it is the analyst who must carry out those inquiries to their logical conclusion.

In the next chapter, we will return to the step we alluded to at the start of this chapter: applying the log-probability techniques set forth previously to the key staffing needs planning problem of estimating hiring needs.

---

[3] Since a single run of GS810 involves the possible operator selection of several different options, each of which will result in a different run format, no one single sample run of GS810 is reproduced here. A reader can look at Appendix F-4 (Sample Outputs) to study the outputs produced by GS810.

CHAPTER 10

STAFFING NEEDS PLANNING COMPUTER PROGRAM: HIREST

First-Year Projections

The final goal of the technology that has been described earlier is to permit personnel management officials to make effective estimates of future staffing needs in the organization(s) they serve. Such future needs estimates are essential if personnel officials are to rationally plan future personnel management programs to meet those needs.

In some circumstances, of course, future needs can be estimated by simple extrapolations of past trends. As we showed earlier, when (a) employment in the future is expected to follow the same trend that it has in the recent past, when (b) the same percentage of accessions are expected to be new outside hires, and when (c) no change is expected in the percentage of the workforce who will be coming to retirement eligibility, then--and only then--current turnover and retirement rates can be expected to continue in the future without significant change. In these cases, simply extend current loss rates into the future unchanged.

If, however, some significant departure from past trend is expected in one or more of these three conditions--employment, percent of new outside hires, or retirement eligibles--then either the turnover rate or the retirement rate, or both, may be expected to change significantly. Where this is so, estimates need to be made of the direction and magnitude of such change(s).

With our final computer program, HIREST, we now show how the L-P techniques described above may be combined with standard actuarial techniques to yield such estimates.

To show how this is done, review Figure 10-1. Note first that the "Growth" portion of hiring needs is determined by the difference between the total population in year X + 1 (next year) and the total population in year X (this year). This difference can be determined by simple subtraction of the starting figure from the target figure.

Note also, however, that by far the bigger share of the hiring needs total is accounted for by "Losses." This is where our combination of techniques is required.

A relatively small proportion of total losses in an average organization consists of losses for actuarial-type reasons: death, disability, and retirement. Figure 10-2 contains a sample of actuarial tables for these

Figure 10-1

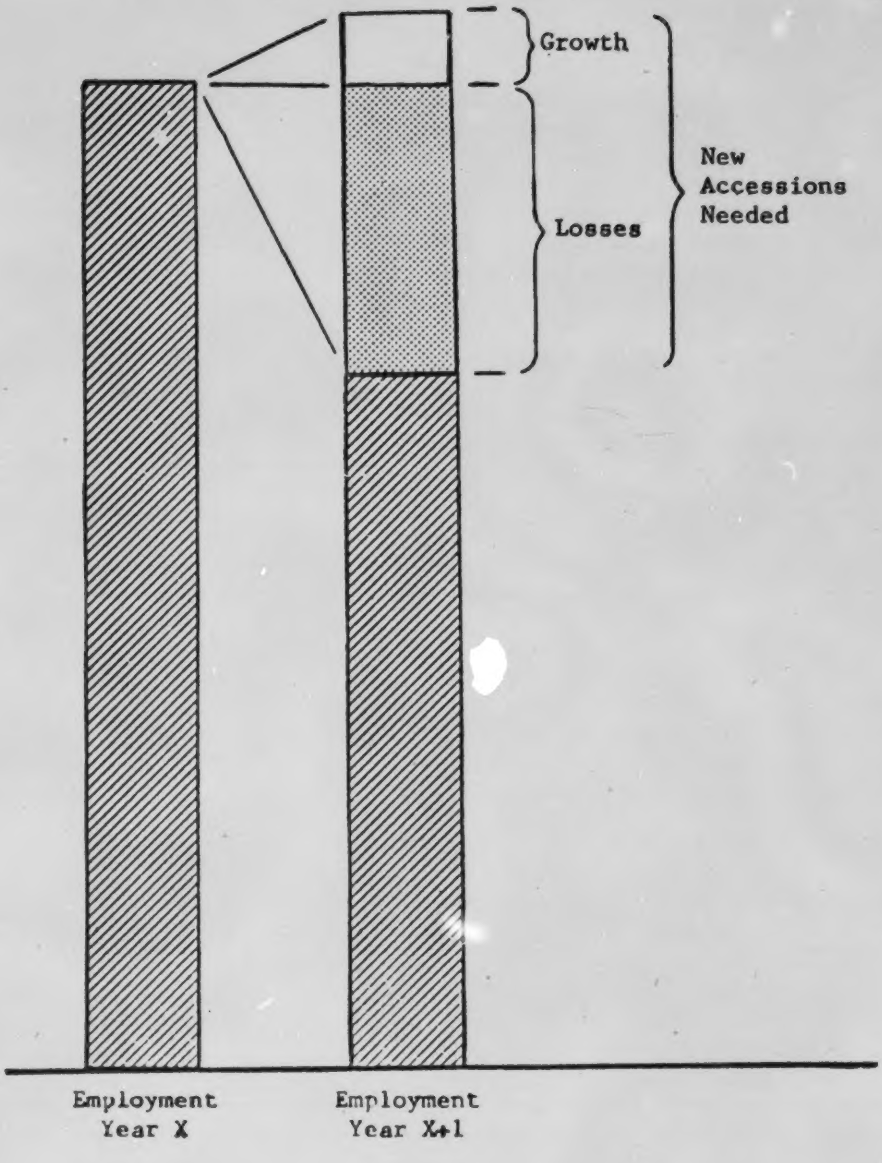

Growth
Losses
New Accessions Needed
Employment Year X
Employment Year X+1

Figure 10-2

SAMPLE ACTUARIAL PROBABILITIES

	MALE PROBABILITIES				FEMALE PROBABILITIES			
			RETIREMENT PROB.				RETIREMENT PROB	
AGE	PROB. OF DEATH	PROB. OF DISAB	50 AND UNDER	51 AND OVER	PROB. OF DEATH	PROB. OF DISAB	50 AND UNDER	51 AND OVER
18.	0.00034	------	-----	-----	0.00032	------	-----	-----
19.	0.00037	------	-----	-----	0.00034	------	-----	-----
20.	0.00040	0.0003	-----	-----	0.00036	0.0006	-----	-----
21.	0.00044	0.0004	-----	-----	0.00039	0.0006	-----	-----
22.	0.00047	0.0004	-----	-----	0.00042	0.0007	-----	-----
23.	0.00052	0.0005	-----	-----	0.00045	0.0008	-----	-----
24.	0.00056	0.0005	-----	-----	0.00048	0.0009	-----	-----
25.	0.00061	0.0006	-----	-----	0.00051	0.0010	-----	-----
26.	0.00067	0.0006	-----	-----	0.00055	0.0011	-----	-----
27.	0.00073	0.0007	-----	-----	0.00059	0.0012	-----	-----
28.	0.00080	0.0008	-----	-----	0.00063	0.0013	-----	-----
29.	0.00087	0.0009	-----	-----	0.00068	0.0014	-----	-----
30.	0.00095	0.0010	-----	-----	0.00073	0.0015	-----	-----
31.	0.00103	0.0011	-----	-----	0.00078	0.0017	-----	-----
32.	0.00112	0.0013	-----	-----	0.00084	0.0019	-----	-----
33.	0.00122	0.0014	-----	-----	0.00090	0.0021	-----	-----
34.	0.00133	0.0016	-----	-----	0.00096	0.0023	-----	-----
35.	0.00145	0.0018	-----	-----	0.00103	0.0025	-----	-----
36.	0.00158	0.0020	-----	-----	0.00110	0.0028	-----	-----
37.	0.00173	0.0022	-----	-----	0.00118	0.0030	-----	-----
38.	0.00188	0.0025	-----	-----	0.00127	0.0034	-----	-----
39.	0.00205	0.0028	-----	-----	0.00136	0.0037	-----	-----
40.	0.00224	0.0031	-----	-----	0.00145	0.0041	-----	-----
41.	0.00244	0.0035	-----	-----	0.00156	0.0045	-----	-----
42.	0.00266	0.0039	-----	-----	0.00167	0.0049	-----	-----
43.	0.00289	0.0043	-----	-----	0.00179	0.0054	-----	-----
44.	0.00316	0.0048	-----	-----	0.00192	0.0060	-----	-----
45.	0.00344	0.0054	-----	-----	0.00206	0.0066	-----	-----
46.	0.00375	0.0061	-----	-----	0.00220	0.0073	-----	-----
47.	0.00408	0.0068	-----	-----	0.00236	0.0080	-----	-----
48.	0.00445	0.0076	-----	-----	0.00253	0.0088	-----	-----
49.	0.00485	0.0085	-----	-----	0.00271	0.0097	-----	-----
50.	0.00529	0.0095	-----	-----	0.00291	0.0107	-----	-----
51.	0.00576	0.0107	-----	-----	0.00311	0.0118	-----	-----
52.	0.00628	0.0119	-----	-----	0.00334	0.0130	-----	-----
53.	0.00685	0.0134	-----	-----	0.00358	0.0143	-----	-----
54.	0.00746	0.0149	-----	-----	0.00383	0.0158	-----	-----
55.	0.00813	0.0167	0.250	-----	0.00411	0.0174	0.290	-----
56.	0.00886	0.0187	0.160	-----	0.00440	0.0191	0.200	-----
57.	0.00966	0.0210	0.150	-----	0.00472	0.0211	0.180	-----
58.	0.01053	0.0235	0.150	-----	0.00506	0.0232	0.180	-----
59.	0.01147	0.0263	0.160	-----	0.00542	0.0256	0.190	-----
60.	0.01251	0.0294	0.210	-----	0.00581	0.0282	0.260	-----
61.	0.01363	0.0329	0.190	-----	0.00622	0.0310	0.200	-----
62.	0.01485	0.0368	0.180	0.060	0.00667	0.0342	0.220	0.080
63.	0.01619	0.0412	0.180	0.050	0.00715	0.0377	0.190	0.060
64.	0.01764	0.0461	0.190	0.060	0.00766	0.0415	0.180	0.070
65.	0.01923	0.0516	0.230	0.110	0.00821	0.0457	0.230	0.120
66.	0.02096	0.0578	0.220	0.090	0.00880	0.0504	0.210	0.100
67.	0.02284	0.0647	0.220	0.090	0.00943	0.0555	0.200	0.100
68.	0.02490	0.0724	0.230	0.100	0.01011	0.0611	0.220	0.130
69.	0.02713	0.0810	0.620	0.100	0.01083	0.0673	0.610	0.140
70.	0.02957	0.0907	1.000	0.110	0.01161	0.0742	1.000	0.150
71.	0.03223	0.1015	-----	0.120	0.01244	0.0817	-----	0.160
72.	0.03513	0.1136	-----	0.130	0.01333	0.0900	-----	0.170
73.	0.03829	0.1271	-----	0.140	0.01429	0.0991	-----	0.180
74.	0.04173	0.1423	-----	1.000	0.01531	0.1092	-----	1.000

three types of losses. Note that for both death and disability the employee data needed are simply "Sex" and "Age." For retirement, however, "Length of service toward retirement" is also needed.

Note also that what the tables provide is the probability of the employee's being lost to that cause during a year. These are, in other words, annual loss probabilities.

We can make up an estimate of the numbers of deaths, disabilities, and retirements to be expected in our workforce in a year, then, by determining each employee's probability of loss under each category and then adding up all of the death probabilities, all of the disability probabilities, and all of the retirement probabilities. The sum of each category will be our estimate of expected losses in that category.

If, for example, we have 100 employees whose retirement probabilities average 0.05, we would expect 5 retirements. If the sum of disability probabilities was 0.95, we would expect one disability. If death probabilities totaled 1.03, we would expect one death.

As you see, if we have each employee's date of birth (DOB), sex, and service computation date (SCD), the finding of probabilities and adding them up to make one-year estimates is simple and straightforward.

To these actuarial estimates, we must next add estimates of the probability of other types of loss from a group during the year. To do this, we will use the L-P equation computed for this group.

Let us assume that the L-P equation for our group--i.e., occupation--is

$$y = 1.1 - 0.9 \log x$$

Then let us assume that we have an employee whose entry on duty (EOD) date is four and one half years prior to the start of our projection year. We first use the equation to get the employee's probability of survival to the start of the projection year by substituting

$$x = 4.5$$

into the equation to get

$$\begin{aligned} y &= 1.1 - 0.9 \log (4.5) \\ y &= 1.1 - 0.9\ (.6532) \\ y &= 1.1 - .5879 = 0.5121 \\ p(y) &= .6957 \end{aligned}$$

Then to get that employee's probability of lasting to the end of the period, we add one year to the 4.5 and substitute

$$x = 5.5$$

into the equation, getting

$$y = 1.1 - 0.9 \log (5.5)$$
$$y = 1.1 - 0.9 (.7404)$$
$$y = 1.1 - .6664 = .4336$$
$$p(y) = .6677$$

To get the probability of the employee's lasting from the start of the year to the end, we divide the ending probability by the starting probability:

$$0.6677/0.6957 = 0.9598$$

Thus, the probability of this employee's being an L-P type loss during the projection year is:

$$1 - 0.9598 = 0.0402$$

By using this method, we can determine the probability of L-P type loss for each employee, using that employee's EOD date. These probabilities can then be added together just like the actuarial loss probabilities to make up an estimate of group L-P type losses for the projection year.

Now, since we can estimate, for each employee, the probability of loss due to death, disability, and retirement reasons by means of actuarial tables, and the probability of loss due to other causes by means of the L-P technique, it is obvious that if we add all of these probabilities together the sum total must equal the total probability of that employee's being lost for all causes during the year. And since we can make such estimates for all employees in the workforce as of year X, and since we can add our individual estimates together to make group estimates, we obviously can make estimates for total group losses during the year from date X to X + 1.

Further Years' Projections

From this point on, however, the mathematics gets a little more complicated because it is not enough to make projections only to point X + 1. We must make them for an additional one (or more) years beyond point X + 1. And to make projections for a second period involves some mathematical problems.

Assume for a moment that you are part way along in a fiscal year. You have just gotten a file of employee data for the workforce on board as of date X, the beginning of the fiscal year which you are now in. If you make a projection of that data through date X plus 1 year, you have made a projection only through the end of the current fiscal year. Obviously, then, if your desire is to make a projection through the Budget Year, or beyond, you will have to make further years' projections in order to get what you need.

The mathematical problem in making such further years' projections does not stem from our basic probability-estimating techniques--these will work perfectly well for any year you name. Rather, the problem stems from the fact that estimates for periods after the first are estimates of conditional probability.

We can explain the problem this way. Suppose you bet a friend that if you flip a coin it will come up heads. What is your probability of winning your bet? Simple enough, 0.50. Suppose you then bet again that you can flip the coin and have it come up heads, what are your chances of winning a second time? Again, 0.50. For each independent trial, that is, the probability of heads is the same. And since each bet was for only one trial, your probability of winning your bet was 0.50 both times.

Now, however, consider the conditional-probability case where what you are betting is that you can flip two heads in a row. Now the question becomes, "What is the probability that you will lose (flip a tails) on the first flip and what is the probability that you will lose (flip a tails) on the second flip?"

The probability of flipping a tails on the first flip is still 0.50, as in the one-trial case. And in those cases where you have gotten a heads on the first flip, the chance of getting a heads on the second flip is again 0.50. But remember, you had a 0.50 probability of getting a tails on the first flip and thus never getting a chance to take a second flip at all. So the probability of your losing on the second flip is 0.50--the probability of getting heads on your second flip--times 0.50, the probability of the first event (getting a heads on the first flip), which must have occurred before the second event (second flip) can take place.

Thus your probability of losing on the first flip is 0.50 and your probability of losing on the second flip is 0.25. Your probability of flipping two heads in a row is therefore what is left, or 0.25.

This multiplying of probabilities together when the probability of a second event is conditional upon the occurrence of a first event also must be done when you are making projections for more than one time period.

Consider, for example, the case of an employee who we determine has an L-P loss probability of 0.10, a death probability of 0.01, a disability probability of 0.01, and a retirement probability of 0.02. Since the employee is present in the workforce at the start of the year, the probability of being present is 1.00, or 100% certainty. We can set up our first-year projection table like this:

147

	Loss Prob.		Prob. of being present, start of year		Net Prob. of loss
L-P loss	0.10	x	1.00	=	0.10
Death	0.01	x	1.00	=	0.01
Disability	0.01	x	1.00	=	0.01
Retirement	0.02	x	1.00	=	0.02
Total					0.14

Probability of retention to end of year:

1.00 - 0.14 = 0.86

For our second projection period, however, we must multiply the rates estimated for each employee who was present at the start of a year by this employee's probability of being present at the start of the second year, like this:

	Loss Prob.		Prob. of being present, start of year		Net Prob. of loss
L-P	0.09	x	0.86	=	0.0774
Death	0.01	x	0.86	=	0.0086
Disability	0.01	x	0.86	=	0.0086
Retirement	0.02	x	0.86	=	0.0172
Total					0.1118

Probability of retention to start of next year:

0.86 - 0.1118 = 0.7482

And so on. Projections for a third or subsequent year would be made in a similar manner, multiplying each annual loss probability by the probability of the employee's being present at the start of the year.

Projecting "Hires Needed"

If you will go back to Figure 10-1 now, you will see that by these methods we have made a projection of the losses which can be expected during the projection period among those employees whom we started out with in year X, the as-of date of our employee data file. And you will note that the sum of "Losses" plus "Growth" equals "Hires," the number of added employees that you need to have on board as of date X + 1 in order for the workforce to be the required size.

Please note very carefully, however, that this number of added employees does not quite equal the total number of new accessions that must be made to the workforce in order to have the needed number of added employees on board at the end of the year. Because your hiring is spread out over the year, some new hires will have quit before the end of the year. So you must actually make more accessions during the year than the number of added employees you need at year's end.

How do we estimate how many more accessions we will need? We use two methods, based on the two different kinds of accessions that can be made, new outside hires, or in-service accessions of employees from other occupations or organizations.

HIREST asks for input on the percentage of total accessions that are new hires (i.e., new Career-Conditional appointments). (This proportion, by the way, tends to be remarkably stable in most cases over even widely-varying conditions.) HIREST then takes this figure--say, 60%--determines the percent filled by in-service accessions--in this case, 40%--and then multiplies these percentages times the "Hires Needed" total from Figure 10-1 to get estimates of the actual numbers of both kinds of accessions who will be needed by date X + 1.[1]

In the first estimating method we use, then, we assume that new hires are spread evenly over the year. Thus, at the end of the year, the newly-hired employees will have an average length of service of 0.5 year. To find what portion of these will have been lost to turnover, we substitute 0.5 into our L-P equation:

$$\begin{aligned} y &= 1.1 - 0.9 \log (0.5) \\ y &= 1.1 - 0.9 (-.30103) \\ y &= 1.1 + 0.27093 = 1.37093 \\ p(y) &= 0.9148 \end{aligned}$$

To find the number of new hires which must be made during the year per employee on board at the end of the year, we divide the retention rate into 1:

$$1/0.9148 = 1.093$$

By multiplying this figure times the number of newly-hired employees needed on board at the end of the year, we get an estimate of the number of new hires we need to make during the year.

Projecting losses from this new-hires group for further years requires only elementary L-P technique. If their average LOS[2] at year X + 1 was 0.5, their average LOS a year later, at year X + 2, is 1.5. Another year later, 2.5. And so on. Substitution of these x-values into the L-P equation gives us a direct estimate of future retention (and by subtraction, of future losses) in the manner described for LOGPRO earlier.

The second estimating method we use is for in-service accessions. In this case, since we cannot estimate their LOS, we must assume that their loss rates will be comparable to those of existing employees. Here, too, however, we must also assume that these accessions will be spread over the whole year so that average service in this occupation is only 0.5 year. Here, therefore, we estimate losses as equal to one-half the annual loss rate for the existing workers. Assuming the existing worker loss rate is 14%, the number of new in-service hires needed per hire on board at the end of the year would be:

---

[1] Note that if the first-six-months loss rate for new hires differs substantially from that for in-service accessions, the final "hires needed" estimates will show a somewhat different ratio of new hires to total accessions than was input into the model (e.g., Figure 10-4).

[2] LOS = Length Of Service

$$1.0 - (0.14 \times 0.5) = 1.0 - 0.07 = 0.93$$
$$1.0/0.93 = 1.075$$

Projecting future losses from this in-service accessions group is also straightforward. Since our starting assumption was that this group's loss rates were similar to those of retained employees, we simply apply our rate estimates for retained employees to this group without change.

HIREST Requirements

Since the above discussion gives a basic picture of HIREST's principal techniques and assumptions, it is appropriate next to describe briefly

a) What inputs HIREST requires;
b) What HIREST does; and
c) What it does not do.

HIREST's input requirements can be summarized briefly. From the above discussion, it will be readily apparent what role each data item plays. Required inputs are:

1) An employee data file showing for each employee in the occupation (in the following order):
    - DOB (Date of Birth)
    - EOD (Entry on Duty) date[3]
    - Sex (1 = F, 2 = M)
    - SCD (Service Computation Date)[4]
2) File "As-of" Date for employee data file
3) The group's L-P loss equation
4) Number of years projections wanted
5) Percentage of total accessions who are new hires (i.e., new Career-Conditional appointments)
6) Population estimates for the years covered

Given these data, HIREST performs its calculations automatically and prints out the results in summary tabulation form, a sample of which is reproduced in Figures 10-3 and 10-4. Note that these tabulations include:

1) A summary of the group population data which was input earlier.

---

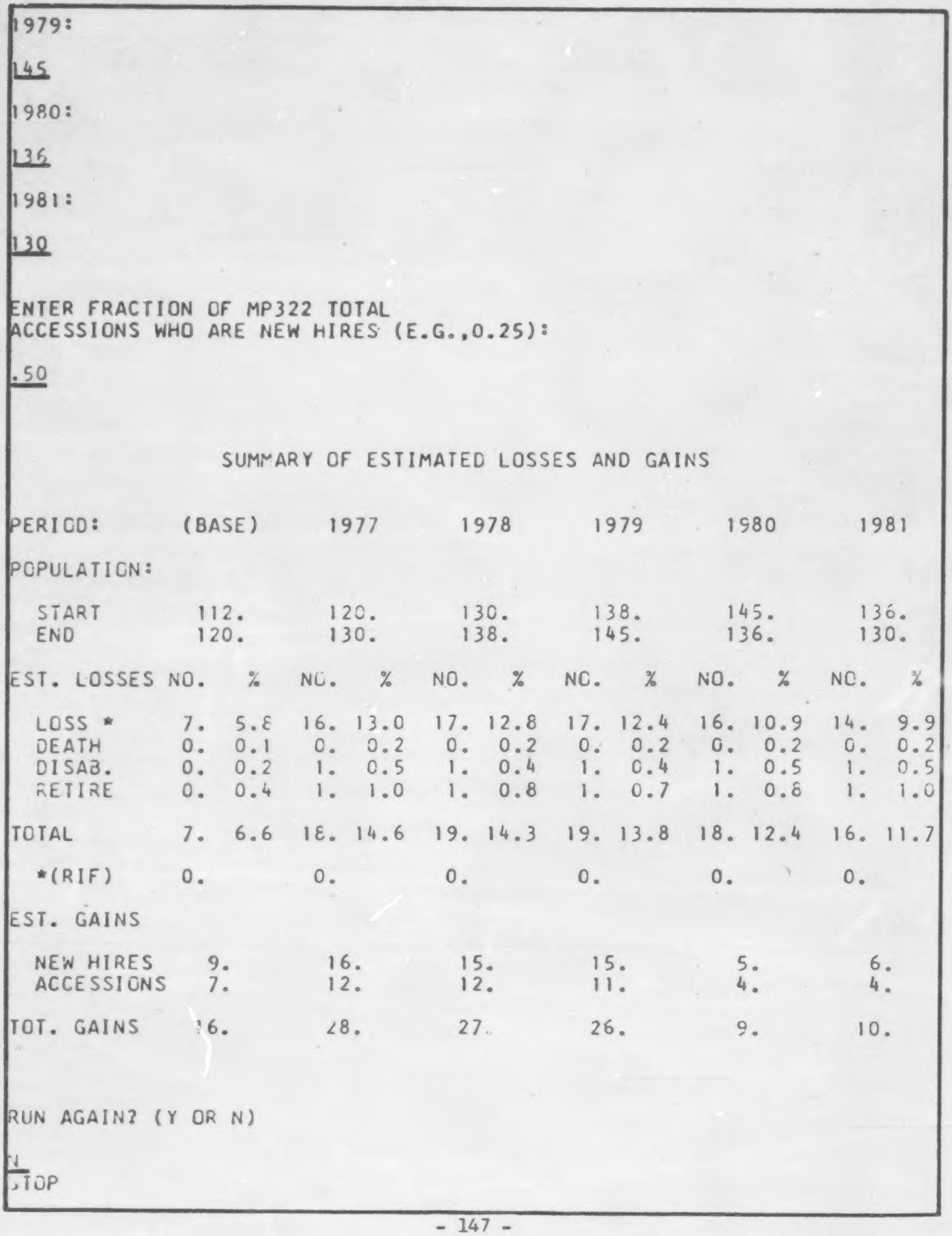
[3] Date of first entry into service: The first entry on the SF 7, Service Record Card, or in the Service Record File, if automated. All dates in month and year only. E.g., 0253 (= 2/53).

[4] A sample line of file for a male born 8/44 with an EOD date of 9/68 and an SCD date of 6/64 would be: 0844096820664.

Figure 10-3

```
ENTER 5-SPACE NAME OF
EMPLOYEE DATA FILE

MP322


ENTER NO. OF EMPLOYEES
IN MP322 FILE


112


ENTER MP322 AS-OF (I.E.,
CURRENT) DATE IN MO., YR. (E.G.,05,75)

1,76


ENTER A,B OF MP322 L-P EQUATION

.4635,-1.35693


FILE MP322 READ. NEXT:


ENTER NO. OF FISCAL YEARS
PROJECTION WANTED (1-5):

5

ENTER MO., YEAR OF START
OF FIRST FISCAL YEAR (E.G., 10,76)

7,76


ENTER MP322 POPULATION
AT START OF FY 1977:

120


ENTER MP322 POPULATION
AT END OF FISCAL YEAR

1977:

130

1978:

138
```

Figure 10-4



```
1979:

145

1980:

136

1981:

130


ENTER FRACTION OF MP322 TOTAL
ACCESSIONS WHO ARE NEW HIRES (E.G.,0.25):

.50



              SUMMARY OF ESTIMATED LOSSES AND GAINS


PERIOD:       (BASE)      1977       1978       1979       1980       1981

POPULATION:

  START        112.       120.       130.       138.       145.       136.
  END          120.       130.       138.       145.       136.       130.

EST. LOSSES NO.   %    NO.   %    NO.   %    NO.   %    NO.   %    NO.   %

  LOSS *     7.  5.8  16. 13.0  17. 12.8  17. 12.4  16. 10.9  14.  9.9
  DEATH      0.  0.1   0.  0.2   0.  0.2   0.  0.2   0.  0.2   0.  0.2
  DISAB.     0.  0.2   1.  0.5   1.  0.4   1.  0.4   1.  0.5   1.  0.5
  RETIRE     0.  0.4   1.  1.0   1.  0.8   1.  0.7   1.  0.8   1.  1.0

TOTAL        7.  6.6  18. 14.6  19. 14.3  19. 13.8  18. 12.4  16. 11.7

  *(RIF)     0.        0.        0.        0.        0.        0.

EST. GAINS

  NEW HIRES   9.        16.       15.       15.        5.        6.
  ACCESSIONS  7.        12.       12.       11.        4.        4.

TOT. GAINS   16.        28.       27.       26.        9.       10.


RUN AGAIN? (Y OR N)

N
STOP
```

2) A summary of loss estimates, by type of loss, with data shown as both number and percent.

3) A special detail line showing the number of losses, if any, which are expected to be RIF's.

4) A summary of gains estimates, broken down by New Hires and In-Service Accessions.

On an overall basis, then, it may be said that what HIREST does is to perform two functions. First, it serves as a projection model by means of which personnel officials are able to make projections of occupational hiring on a multi-year basis.

Second, it serves as a simulation model whose input variables can be deliberately varied to test the effects of input changes on occupational hiring levels, turnover levels, etc. One can test the effects of specific changes in population levels, for example, on the number of turnover losses, or RIF's, or in-service accessions. In this way, HIREST can help advise management of some of the personnel implications of proposed management actions before such actions are taken.

It should also be noted, however, that there are certain things that HIREST is not or does not do. First, it is not infallible. It provides a means of making "probable value" estimates. But these are only estimates. They are not last-digit-accuracy predictions. There are too many uncertainties in the total process to make unqualified predictions a realistic possibility. As one example, HIREST assumes that hiring will be evenly distributed over each of the projection years and accordingly has estimated turnover among new hires using a 0.5 averaging factor. If real hiring is concentrated early in the year (averaging factor, say, 0.7), however, or late in the year (factor of, say, 0.3), HIREST estimates will be significantly off target as a result.

Second, and in a way related to the first point, HIREST is a stochastic or probabilistic model but it does not provide confidence-range estimates for its projections. In the program LOGPRO, for example, we provide limits above and below projected values within which actual values can be expected to fall in 95% of possible cases. In HIREST, such 95%-confidence limits are not provided.

In part, this reflects the situation discussed first, above, that there are many unknown variables whose effects cannot be estimated beforehand. And in part, this reflects the technical problem that whereas we do have confidence-level data on L-P curve equations, we do not have such data for the actuarial tables that make up a substantial part of our loss estimates. For both reasons, providing 95%-confidence limits for HIREST projections is not feasible.

And third, while HIREST does provide estimates of both in-service and outside hiring, it does not provide any breakdown of hires by grade level. These will have to be estimated by the analyst by other means.

Conclusions

HIREST provides a very powerful and very flexible tool for staffing needs analysis. Combining two techniques of proven validity--actuarial technique and log-probability analysis--it enables personnel officials (a) to simulate for management the major turnover and hiring implications of management's alternative workforce plans, and (b) to project the future turnover and hiring levels which may be expected under management's approved workforce plan. Thus, HIREST performs the central analytical tasks necessary for staffing needs planning in the operational setting.

As with any projection model, of course, HIREST projections are only probable values--they are in no sense predictions possessing any last-digit accuracy. Also, HIREST projections reflect the basic assumptions used in the model, such as hiring distributed evenly during projection years. To the extent that these assumptions are not borne out, HIREST results may be affected accordingly.

On the whole, however, HIREST can generate a great deal of very valuable management and personnel management information from relatively simple and straightforward elements. We believe that it is an effective first step toward the still more powerful and refined models of the future.

CHAPTER 11

THE ROLE OF THE ANALYST

When Abraham Lincoln delivered his famous "House Divided" speech in 1858, it was at a time of unprecedented danger for the Union. The United States would soon be engaged in a great civil war and the responsibilities facing the Nation's leaders were awesome and forbidding. In the very opening lines of this speech, which is remembered as his most notable address before becoming President, he summed up the information needs of our leaders at that fateful hour in words of matchless precision:

> "If we could first know where we are, and whither we are tending, we could better judge what to do, and how to do it...."

Every analyst should memorize these words. Not only because they express with absolute clarity the information needs of every decision-maker. But also because they show the way for every analyst, no matter where located, to make a vital contribution to the effectiveness of our system of government.

You have seen in our earlier discussions how to perform the key analytic functions of staffing needs planning. You have seen the policies and functional provisions needed for effective staffing needs planning programs. You know what planning data are needed from management. You know how to analyze and project turnover. You know how to estimate future staffing needs.

With this knowledge, you can now provide decision-makers with many types of information which they vitally need to do their jobs better. If the manager proposes a future course of action requiring a workforce structure that cannot be staffed, you can now say so. If there are workforce plan changes which could be made and which would make staffing feasible, you can make known those needed changes. If the needed workforce cannot be delivered at the cost specified, you can provide better estimates. And, if the required workforce cannot be delivered on the schedule necessary under management's program plan, you can inform the manager of this.

These are information items that the manager urgently needs to be sure that the plans being made, or the alternative program proposals being weighed, are in fact real and viable--plans which in actual fact can be carried out.

If, because these data were not provided, the manager submits to top executives, or to legislative bodies, program plans or decision alternatives which in actual fact:

155

- cannot be carried out at all because the needed personnel are not available; or which
- cannot be carried out within the cost levels specified; or which
- cannot be carried out on schedule because needed hiring and/or training cannot be completed in time;

then because of the failure of the analyst, executive and legislative actions will be taken and program performance promises will be made that the passage of time will show were seriously in error.

In such circumstances, the work of executive and legislative decision-makers can become no more than guesswork. And the performance promises of government can become literally incredible to the very public that government exists to serve.

Such failures of analysis as these can thus contribute in no small measure to the severe impairment of public confidence in the word and workings of government.

Clearly, as we said in the beginning, major improvements in analytical methods and techniques are urgently and vitally needed. And the further we progress toward truly multi-year planning, the more intense this need becomes.

Everyone's contribution is needed in this improvement effort. The techniques described in this handbook, we believe, represent one major step forward. But they are not by any means the last possible word.

On the contrary, they will in their turn be supplemented and eventually supplanted by other, still more effective techniques. Perhaps some of the readers of this handbook may be given or may take the opportunity to contribute to this progress. If you do get such an opportunity, we hope you will take it and that you will give it your best effort.

Every improvement that can be made in staffing needs planning--or indeed in the whole manpower planning process--is an improvement in the ability of government managers--and thus, of our executive and legislative officials--to make public decisions more effectively.

Every improvement in analysis, then, enhances the ability of our governmental system to be responsive to the informed wishes and choices of the people at large.

And that, ultimately, is what the job of the analyst--and indeed that of every public servant--is all about.

APPENDICES

APPENDIX A

MANUAL CALCULATION OF THE LOG-PROBABILITY CURVE

MANUAL CALCULATION OF THE LOG-PROBABILITY CURVE

Problem:

Determine the A and B values of a log-probability retention equation using:

- Manual least squares techniques
  and
- Normal curve area and logarithm conversion tables.

Data Collection:

Retention data are obtained by determining the number or percent of employees retained from a given group of hires after specified lengths of service. Ideally, a group should be composed of employees in the same occupation. However, if the number of employees in one or more occupations is small, then it is possible to combine like occupations.

To analyze a given group of employees, two types of information must be recorded:

(1) The number of employees in the original group of hires; and

(2) The number (or percent) of these employees retained after given lengths of time.

The "original group" is composed of employees hired during a specific time span (one year or less). The "given lengths of time" consist of a standard time unit and, in most cases, an averaging factor which approximates the actual average length of service of the original group at the end of the hiring span.

Sample Data:

The following is a sample set of retention data which will be used in the calculations below:

Number hired in occupation X in Fiscal Year 1970 = 250

At end of FY:	Number Retained	Percent Retained
71	208	83.2%
72	189	75.6%
73	176	70.4%
74	164	65.6%

The averaging factor used will be 0.5 years (which assumes an even distribution of hiring during a fiscal year). Thus, the value for FY 71 is 1.5 years, for FY 72 is 2.5, etc.

The Equation:

The mathematical form of a line is given by the equation:

$$Y = A + BX$$

Retention data (also called log-probability data) can be expressed linearly by making two transformations:

(1) Percent of employees retained is transformed into number of standard deviations from the mean of a normal curve; and

(2) Years of service completed is transformed into the logarithm of years of service completed.

These two transformations can be made using the two tables included in this instruction:

(1) The Normal Curve Area Conversion Table

(2) The Table of Logarithms

Computations:

To solve the stated problem it is necessary to transform the retention data into their linear form and solve for the coefficients A and B.

To begin, let:

N = Number of retention points (i.e., number of observations)

i = Retention point i, $i=1,2,\ldots,N$

t_i = Length of service represented by retention point i

X_i = Base 10 logarithm of t_i

$P(t_i)$ = Percent of the original group remaining at time t_i

Y_i = Number of standard deviations from the mean of a normal curve represented by the value $P(t_i)$

This instruction will compute A and B using the sample data and:

Case I: Two years of retention data

Case II: Three or more years of retention data

Case I. If only two years of retention data (or two retention points) are available, then the A and B values of the log-probability equation may be computed using the following formulas:

$$B = \frac{Y_1 - Y_2}{X_1 - X_2} \qquad (1)$$

$$A = Y_1 - BX_1 \qquad (2)$$

Using the first two points of the sample data:

$t_1 = 1.5 \quad P(t_1) = 83.2\%$
$t_2 = 2.5 \quad P(t_2) = 75.6\%$

From the Table of Logarithms, we obtain:

$X_1 = 0.17609$
$X_2 = 0.39794$

And from the Normal Curve Area Conversion Table and interpolation (see Page 7):

$Y_1 = 0.96223$
$Y_2 = 0.69358$

Substituting these values into formulas (1) and (2), we get:

$$B = \frac{0.96223 - 0.69358}{0.17609 - 0.39794}$$

$$= -1.210953$$

And,

$$A = 0.96223 - (-1.210953)(0.17609)$$

$$= 1.175467$$

Thus, the log-probability retention equation in this case is:

$$Y = 1.175467 - 1.210953X$$

Case II. If there are three or more years of retention data available, the A and B values of the log-probability equation may be calculated in one of two ways:

(1) Simultaneous solution of the "normal equations"; or

(2) Substitution of values into the "linear fit algorithm".

These methods are both variations of the "least squares" technique.

Whichever method is chosen, several quantities must be calculated. For this purpose, it is useful to set up a work table in the following form (using all the sample data points):

Year t_i	Log Year* X_i	Percent Remaining $P(t_i)$	Stan. Dev. From Mean** Y_i	$X_i x Y_i$	X_i^2
1.5	0.17609	83.2	0.96223	0.16944	0.03101
2.5	0.39794	75.6	0.69358	0.27600	0.15836
3.5	0.54407	70.4	0.53599	0.29162	0.29601
4.5	0.65321	65.6	0.40160	0.26233	0.42668
	1.77131		2.59340	0.99939	0.91206

* Using Table of Logarithms

** Using Normal Curve Area Conversion Table and interpolation

II-1. The normal equations are:

$$\Sigma Y = NA + B\Sigma X \quad (3)$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2 \quad (4)$$

From the above work table, we have:

$$\Sigma X = 1.77131$$
$$\Sigma Y = 2.59340$$
$$\Sigma XY = 0.99939$$
$$\Sigma X^2 = 0.91206$$
$$N = 4$$

Substituting these values into equations (3) and (4), we obtain:

$$2.59340 = 4A + 1.77131B \quad (3a)$$
$$0.99939 = 1.77131A + 0.91206B \quad (4a)$$

Solving these equations simultaneously (see Page 8), we have:

$$A = 1.165275$$

$$B = -1.167328$$

And the log-probability equation is:

$$Y = 1.165275 - 1.167328X$$

II-2. The linear fit algorithm is:

$$B = \frac{N\Sigma XY - \Sigma X\Sigma Y}{N\Sigma X^2 - (\Sigma X)^2} \quad (5)$$

$$A = \frac{\Sigma Y - B\Sigma X}{N} \quad (6)$$

Using the values from the work table, we get:

$$B = \frac{4(0.99939) - (1.77131)(2.59340)}{4(0.91206) - (1.77131)^2}$$

$$= \frac{-0.596155}{0.510701}$$

$$= -1.167328$$

And,

$$A = \frac{2.59340 - (-1.167328)(1.77131)}{4}$$

$$= 1.165275$$

Thus, the log-probability equation is:

$$Y = 1.165275 - 1.167328X \qquad (7)$$

Iteration and Projection:

The processes of iteration and projection are conceptually the same. They both involve the substitution of X-values into a given equation to obtain the Y-values given by the line (Y'). The only difference between the two is in the X-values which are used.

In iteration, previously-observed (or past) X-values are used to obtain the equation values. For example, using equation (7) and the first X-value in the sample data (i.e., $X_1 = 0.17609$), we obtain:

$$Y'_1 = 1.165275 - 1.167328(0.17609)$$
$$= 0.95972$$

Using the Normal Curve Area Conversion Table and interpolation, this value may be converted to percent. In this case, the percent value is 83.14%.

Iterated values are used in the calculation of standard deviation and variance and other goodness of fit measures.

For projection, future X-values are used to obtain estimates of the number retained from an original group at future points in time. For example, using equation (7) and year 5.5, we have:

$$t = 5.5$$

$$X = 0.74036$$

And,

$$Y' = 1.165275 - 1.167328(0.74036)$$
$$= 0.30103 \quad \text{(or 61.83\%)}$$

Projected values may be used in planning for future hiring needs.

Standard Deviation:

To calculate a standard deviation from the log-probability line of regression it is necessary to (a) iterate the regression equation for all the past values of X and (b) determine the differences between the actual Y-values and the curve values ($Y^{\sim}$). This can be done using the following work table and the sample data:

Year (t_i)	Actual (Y_i)	Curve ($Y_i^{\sim}$)	Diff. ($D=Y_i-Y_i^{\sim}$)	Diff.^{2} (D^2)
1.5	0.96223	0.95972	0.00251	0.0000063
2.5	0.69358	0.70075	-0.00717	0.0000514
3.5	0.53599	0.53017	0.00582	0.0000339
4.5	0.40160	0.40276	-0.00116	0.0000013

The formula for the standard deviation is:

$$s = \sqrt{\frac{\Sigma D^2}{(N-1)}}$$

In this example,

$$\Sigma D^2 = 0.0000929$$

$$N = 4$$

And,

$$s = \sqrt{\frac{0.0000929}{(4-1)}} = 0.005565$$

Note that this value is in standard deviations from the mean of the normal curve.

The variance equals the square of the standard deviation.

Confidence Interval for Projections:

In addition to simply projecting a log-probability equation for future points, it is also possible to calculate a 95% confidence range for each projected point. This is done using the standard deviation. To begin, let:

s = Standard deviation from the log-probability line

$Y^{\sim}$ = A projected curve value

And,

$$H = Y^{-} + 2s$$

$$L = Y^{-} - 2s$$

Both H and L are in the form standard deviations from the mean of the normal curve. Using the Normal Curve Area Conversion Table, these values can be converted to percents and these percents will constitute the confidence range. To convert to numerical values, simply multiply each percent (in decimal form) by the number in the original starting group.

For example, using equation (7) and $t = 5.5$ years, then:

$$Y^{-} = 0.30103$$

And,

$$\begin{aligned} s &= 0.005565 \\ 2s &= 0.011130 \end{aligned}$$

Thus,

$$H = 0.30103 + 0.01113 = 0.31216$$

$$L = 0.30103 - 0.01113 = 0.28990$$

Using the Normal Curve Area Conversion Table and interpolation, H and L convert to 62.25% and 61.41%, respectively. The percent value associated with Y^{-} is 61.83% (this is also known as the "expected value"). Numerically, the range would be (with 250 in the starting group):

High: .6225 x 250 = 155.6

Expected Value: .6183 x 250 = 154.6

Low: .6141 x 250 = 153.5

For planning purposes, these figures may be rounded.

Linear Interpolation:

The process of linear interpolation is used to read between the lines of a statistical table; i.e., to calculate intermediate values.

In general, suppose that value V_1 is represented in a table by value T_1 and V_2 is represented by T_2. What is the T-value associated with the value V ($V_2<V<V_1$)? The value of T can be calculated from the following ratio:

$$\frac{V - V_2}{V_1 - V_2} = \frac{T - T_2}{T_1 - T_2}$$

Solving this equation for T, we get:

$$T = T_2 + \left(\frac{V - V_2}{V_1 - V_2}\right)(T_1 - T_2) \qquad (8)$$

This process can be used in both of the tables attached to this instruction.

For example, using the sample data and the Normal Curve Area Conversion Table, what is the standard deviation value associated with 83.2%? In this case,

$$\begin{array}{ll} V = 83.2\% & T = ? \\ V_1 = 84.0\% & T_1 = 0.99446 \\ V_2 = 83.0\% & T_2 = 0.95417 \end{array}$$

Using equation (8).

$$T = 0.95417 + \left(\frac{83.2 - 83.0}{84.0 - 83.0}\right)(0.99446 - 0.95417)$$

$$= 0.95417 + (0.2)(0.04029)$$

$$= 0.96223$$

Thus, the table value associated with 83.2% is 0.96223.

Interpolation may also be used in the other direction; i.e., with the standard deviations columns representing the V-values and the percent remaining columns the T-values.

For example, the first iterated value from the log-probability equation is 0.95972. With what percent value is this associated? In this case,

$$\begin{array}{ll} V = 0.95972 & T = ? \\ V_1 = 0.99446 & T_1 = 84.0\% \\ V_2 = 0.95417 & T_2 = 83.0\% \end{array}$$

And,

$$T = 83.0 + \left(\frac{0.95972 - 0.95417}{0.99446 - 0.95417}\right)(84.0 - 83.0)$$

$$= 83.0 + 0.14$$

$$= 83.14\%$$

Simultaneous Equations:

Solving the normal equations for A and B (see Page 4) is equivalent to solving two simultaneous equations in two unknowns. These equations are of the form:

$$S_1 = Q_1A + Q_2B \qquad (9)$$

$$S_2 = R_1A + R_2B \qquad (10)$$

where S_1, S_2, Q_1, Q_2, R_1, and R_2 are all known quantities.

The technique for solving two simultaneous equations in two unknowns can be shown using the normal equations in this instruction as an example. These equations are:

$$2.59340 = 4A + 1.77131B \qquad (3a)$$
$$0.99939 = 1.77131A + 0.91206B \qquad (4a)$$

The first step in solving these equations is to eliminate one of the unknown values (A or B) from them. If A is selected for elimination, then each term in equation (3a) is multiplied by the coefficient of A in equation (4a) (i.e., 1.77131) and each term in equation (4a) is multiplied by the coefficient of A in equation (3a) (i.e., 4). After these multiplications have been completed the two normal equations now look like this:

$$4.593715 = 7.085240A + 3.137539B \qquad (3b)$$
$$3.997560 = 7.085240A + 3.648240B \qquad (4b)$$

The next step is to subtract one of these equations from the other to get one equation containing only one unknown (since the A terms will cancel out). If equation (4b) is subtracted from equation (3b) the result is:

$$0.596155 = -0.510701B$$

Solving this equation for B gives:

$$B = \frac{0.596155}{-0.510701}$$

$$= -1.167328$$

This calculated value of B can now be substituted back into one of the two original equations to get a value for A. If equation (3a) is selected, the result would be:

$$2.59340 = 4A + 1.77131(-1.167328)$$

And,

$$A = 1.165275$$

It is possible to check these results by substituting the calculated values for A and B into equation (4a). This would give:

$$\begin{aligned} 0.99939 &= 1.77131(1.165275) + 0.91206(-1.167328) \\ &= 2.064063 - 1.064673 \\ &= 0.99939 \end{aligned}$$

TABLE OF LOGARITHMS

YEAR	LOG (10)	YEAR	LOG (10)	YEAR	LOG (10)
0.1	-1.00000	5.1	0.70757	10.1	1.00432
0.2	-0.69897	5.2	0.71600	10.2	1.00860
0.3	-0.52288	5.3	0.72428	10.3	1.01284
0.4	-0.39794	5.4	0.73239	10.4	1.01703
0.5	-0.30103	5.5	0.74036	10.5	1.02119
0.6	-0.22185	5.6	0.74819	10.6	1.02531
0.7	-0.15490	5.7	0.75587	10.7	1.02938
0.8	-0.09691	5.8	0.76343	10.8	1.03342
0.9	-0.04576	5.9	0.77085	10.9	1.03743
1.0	0.00000	6.0	0.77815	11.0	1.04139
1.1	0.04139	6.1	0.78533	11.1	1.04532
1.2	0.07918	6.2	0.79239	11.2	1.04922
1.3	0.11394	6.3	0.79934	11.3	1.05308
1.4	0.14613	6.4	0.80618	11.4	1.05690
1.5	0.17609	6.5	0.81291	11.5	1.06070
1.6	0.20412	6.6	0.81954	11.6	1.06446
1.7	0.23045	6.7	0.82607	11.7	1.06819
1.8	0.25527	6.8	0.83251	11.8	1.07188
1.9	0.27875	6.9	0.83885	11.9	1.07555
2.0	0.30103	7.0	0.84510	12.0	1.07918
2.1	0.32222	7.1	0.85126	12.1	1.08279
2.2	0.34242	7.2	0.85733	12.2	1.08636
2.3	0.36173	7.3	0.86332	12.3	1.08991
2.4	0.38021	7.4	0.86923	12.4	1.09342
2.5	0.39794	7.5	0.87506	12.5	1.09691
2.6	0.41497	7.6	0.88081	12.6	1.10037
2.7	0.43136	7.7	0.88649	12.7	1.10380
2.8	0.44716	7.8	0.89209	12.8	1.10721
2.9	0.46240	7.9	0.89763	12.9	1.11059
3.0	0.47712	8.0	0.90309	13.0	1.11394
3.1	0.49136	8.1	0.90849	13.1	1.11727
3.2	0.50515	8.2	0.91381	13.2	1.12057
3.3	0.51851	8.3	0.91908	13.3	1.12385
3.4	0.53148	8.4	0.92428	13.4	1.12710
3.5	0.54407	8.5	0.92942	13.5	1.13033
3.6	0.55630	8.6	0.93450	13.6	1.13354
3.7	0.56820	8.7	0.93952	13.7	1.13672
3.8	0.57978	8.8	0.94448	13.8	1.13988
3.9	0.59106	8.9	0.94939	13.9	1.14301
4.0	0.60206	9.0	0.95424	14.0	1.14613
4.1	0.61278	9.1	0.95904	14.1	1.14922
4.2	0.62325	9.2	0.96379	14.2	1.15229
4.3	0.63347	9.3	0.96848	14.3	1.15534
4.4	0.64345	9.4	0.97313	14.4	1.15836
4.5	0.65321	9.5	0.97772	14.5	1.16137
4.6	0.66276	9.6	0.98227	14.6	1.16435
4.7	0.67210	9.7	0.98677	14.7	1.16732
4.8	0.68124	9.8	0.99123	14.8	1.17026
4.9	0.69020	9.9	0.99564	14.9	1.17319
5.0	0.69897	10.0	1.00000	15.0	1.17609

NORMAL CURVE AREA CONVERSION TABLE

PERCENT REMAINING	STANDARD DEVIATIONS FROM MEAN	PERCENT REMAINING	STANDARD DEVIATIONS FROM MEAN	PERCENT REMAINING	STANDARD DEVIATIONS FROM MEAN
95	1.64485	65	0.38532	35	-0.38532
94	1.55477	64	0.35846	34	-0.41246
93	1.47579	63	0.33185	33	-0.43991
92	1.40507	62	0.30548	32	-0.46770
91	1.34076	61	0.27932	31	-0.49585
90	1.28155	60	0.25335	30	-0.52440
89	1.22653	59	0.22755	29	-0.55338
88	1.17499	58	0.20189	28	-0.58284
87	1.12639	57	0.17637	27	-0.61281
86	1.08032	56	0.15097	26	-0.64335
85	1.03643	55	0.12566	25	-0.67449
84	0.99446	54	0.10043	24	-0.70630
83	0.95417	53	0.07527	23	-0.73885
82	0.91537	52	0.05015	22	-0.77219
81	0.87790	51	0.02507	21	-0.80642
80	0.84162	50	0.00000	20	-0.84162
79	0.80642	49	-0.02507	19	-0.87790
78	0.77219	48	-0.05015	18	-0.91537
77	0.73885	47	-0.07527	17	-0.95417
76	0.70630	46	-0.10043	16	-0.99446
75	0.67449	45	-0.12566	15	-1.03643
74	0.64335	44	-0.15097	14	-1.08032
73	0.61281	43	-0.17637	13	-1.12639
72	0.58284	42	-0.20189	12	-1.17499
71	0.55338	41	-0.22754	11	-1.22653
70	0.52440	40	-0.25335	10	-1.28155
69	0.49585	39	-0.27932	9	-1.34076
68	0.46770	38	-0.30548	8	-1.40507
67	0.43991	37	-0.33185	7	-1.47579
66	0.41246	36	-0.35846	6	-1.55477
				5	-1.64485

APPENDIX B

SOME APPLICATIONS OF THE LOG-PROBABILITY EQUATION

Question: What is the probability that an employee who has already had
1 t years of Federal service will have at least one more?

Suppose that the log-probability equation for a given employee's occupation is:

$$Y = 0.9534 - 0.9102\text{Log}(t)$$

And suppose that,

t = the employee's length of service (preferably based on original entry on duty date)
= 3.75 years

$t+1$ = the employee's length of service plus one year
= 4.75 years

Then,

$$\begin{aligned} Y_t &= 0.9534 - 0.9102\text{Log}(t) \\ &= 0.9534 - 0.9102\text{Log}(3.75) \\ &= 0.9534 - 0.9102(0.57403) \\ &= 0.43092 \end{aligned}$$

Converting Y_t (which is in standard deviation units) to percent gives:

$$P_t = 0.6667$$

And,

$$\begin{aligned} Y_{t+1} &= 0.9534 - 0.9102\text{Log}(t+1) \\ &= 0.9534 - 0.9102\text{Log}(4.75) \\ &= 0.9534 - 0.9102(0.67669) \\ &= 0.33748 \end{aligned}$$

Converting Y_{t+1} to percent gives:

$$P_{t+1} = 0.6321$$

The retention probability is given by the formula:

$$R = \frac{P_{t+1}}{P_t}$$

In this case,

$$R = \frac{.6321}{.6667} = 0.9481$$

Thus, the probability that an employee who has served 3.75 years will serve 4.75 years is 0.9481.

Question: 2 What is the probability that an accession to your organization will stay with you for at least two years if (a) the accession is a new outside hire and (b) the accession is a reassigned employee who has already been with your organization for 3 years?

Suppose that the log-probability equation for the occupation in which a vacancy is to be filled is:

$$Y = 0.9534 - 0.9102\text{Log}(t)$$

Case a. Suppose that this vacancy is filled by a new outside hire. The probability that a new hire will be with the organization in two years is found by iterating the log-probability equation for the t-value of 2. This gives:

$$\begin{aligned} Y &= 0.9534 - 0.9102\text{Log}(2) \\ &= 0.9534 - 0.9102(0.30103) \\ &= 0.67940 \end{aligned}$$

This Y-value transforms to a probability value of 0.7516.

Thus, for this occupation, there is a 75% chance that a new hire will stay with the job for at least 2 years.

Case b. Suppose that the vacancy is filled by a reassigned employee who has already been with the organization for 3 years. The probability that this employee will stay for at least two more years is found by calculating the quotient of the probability of staying 5 (3+2) years and the probability of staying 3 years. This involves two iterations of the log-probability equation:

$$\begin{aligned} Y_{+3} &= 0.9534 - 0.9102\text{Log}(3) \\ &= 0.9534 - 0.9102(0.47712) \\ &= 0.51913 \end{aligned}$$

This converts to a probability value of:

$$P_{+3} = 0.6982$$

And,

$$\begin{aligned} Y_{+5} &= 0.9534 - 0.9102\text{Log}(5) \\ &= 0.9534 - 0.9102(0.69897) \\ &= 0.31720 \end{aligned}$$

This converts to a probability value of:

$$P_{+5} = 0.6245$$

Then the probability that an employee with 3 years of service will stay at least 5 years is:

$$\frac{P_{+5}}{P_{+3}}$$

Using the calculated probability values gives:

$$\frac{0.6245}{0.6982} = 0.8944$$

Thus, for this occupation, there is an 89% chance that an employee with 3 years of service will stay 2 more. In other words, 9 times out of 10 a selection of someone with 3 years of service will result in that employee staying with the new job for at least 2 years.

The important point to note from this example is the difference in retention probabilities which is caused solely by different methods of selection. In this example we used a log-probability equation which characterizes a professional occupation. However, the differences noted here are even more pronounced in clerical occupations where there can be a 10 to 1 retention difference between selecting an employee already on board or hiring from the outside.

Question: How can individual retention probabilities for employees in
3 a given group be translated into the estimated number who will be lost from the group?

Suppose that the log-probability equation for the given group is:

$$Y = 0.9534 - 0.9102\text{Log}(t)$$

And suppose that the given group consists of 10 employees whose length of service composition and individual retention probabilities (see previous question) are given by the following table:

	-Base Year-		Base year + 1		
Employee Number (i)	L.O.S. (t)	Ret % (P_t)	L.O.S. (t+1)	Ret % (P_{t+1})	Retention Prob. ($R_i = P_{t+1}/P_t$)
1	1.2	.8109	2.2	.7395	.9119
2	3.5	.6766	4.5	.6401	.9461
3	0.8	.8512	1.8	.7646	.8983
4	2.4	.7282	3.4	.6807	.9348
5	1.8	.7646	2.8	.7076	.9255
6	4.3	.6468	5.3	.6157	.9519
7	6.7	.5798	7.7	.5582	.9627
8	5.0	.6245	6.0	.5968	.9556
9	2.1	.7454	3.1	.6936	.9305
10	1.6	.7786	2.6	.7176	.9217

The number lost from this group from (base year) to (base year + 1) is given by:

$$L = N - \sum_{i=1}^{N} R_i$$

From the above table,

$$N = \text{the number of employees in the group} = 10$$

$$\sum_{i=1}^{N} R_i = 9.339$$

And,

$$L = 10 - 9.339 = 0.661$$

This value would round to 1. Thus, 1 employee will be lost from the group. There is no way to determine which one it will be.

Question 4: How many civil engineers must be hired by the organization in this fiscal year in order to have 100 of them on board five years hence?

Assume that the log-probability equation for civil engineers is:

$$Y = 0.9534 - 0.9102\text{Log}(t) \qquad (1)$$

The t-value for base year + 5 is 5.5 years (assuming an even distribution of hires over a fiscal year).

The first step is to calculate the "percent remaining" value given by equation (1) for 5.5 years. Solving the equation for the Y-value associated with t = 5.5 results in:

$$\begin{aligned} Y_{+5} &= 0.9534 - 0.9102\,\text{Log}(t) \\ &= 0.9534 - 0.9102(0.74036) \\ &= 0.27952 \end{aligned}$$

This value is in standard deviation units. Converting to percent gives:

$$P_{+5} = 0.6101 \quad \text{(or 61.01\%)}$$

Thus 61.01% of the starting group hired during the current fiscal year will still be on board in five years.

To determine the number of civil engineers to be hired this fiscal year, let:

N_0 = the number to be hired in the base year

N_{+5} = the number wanted on board in five years

Then,

$$\begin{aligned} N_0 &= \frac{N_{+5}}{P_{+5}} \\ &= \frac{100}{.6101} = 164 \end{aligned}$$

Thus, 164 civil engineers must be hired this fiscal year so that 100 of them will still be on board in five years.

175

Question: What is the three-year return on a training investment (A) when only new outside hires are trained and (B) when employees who have already been on board for three years are trained?
5

Suppose that an organization wishes to train 100 personnel management specialists and that the billed cost of the training program is $1000 per employee trained. And suppose that the log-probability equation for personnel management specialists is:

$$Y = 1.165 - 1.167\text{Log}(t) \qquad (1)$$

Each of the cases described in the question can be explained by calculating two figures:

1. The number of employees from the starting group of 100 who will be retained by the organization three years after the completion of the training course; and
2. The actual cost of the training program per retained employee.

Case A - Suppose that all 100 employees in the training program are selected from a group of new outside hires.

Three years after the training program, those employees still on board will have an average length of service of 3.5 years (assuming an even distribution of hiring during a year).

The Y-value associated with this value of t (i.e., 3.5 years) is given by:

$$\begin{aligned} Y_{+3} &= 1.165 - 1.167\text{Log}(3.5) \\ &= 1.165 - 1.167(0.54407) \\ &= 0.53007 \end{aligned}$$

This value is in standard deviation units. Converting it to percent results in:

$$P_{+3} = 0.7020 \quad \text{(or 70.20\%)}$$

Thus, the number of employees remaining from the original group of 100 is:

$$\begin{aligned} N_{+3} &= P_{+3} \times 100 \\ &= 0.7020 \times 100 = 70 \end{aligned}$$

To translate this number into an actual cost per employee trained figure, let:

TC = the original stated cost per employee of the training

AC = the actual cost of the training program per employee still on board three years later

Then,

$$AC = \frac{TC}{P_{+3}}$$

$$= \frac{\$1000}{.7020} = \$1425/\text{employee trained}$$

Thus, the three-year return on the training investment when only new outside hires are trained is $1425 per employee trained and still on board.

Case B - Suppose that all 100 employees in the training program are employees who have already served three years with the organization.

Three years after training, those trained employees still with the organization will have an average length of service of 6.5 years. The Y-value associated with this value of t is given by:

$$\begin{aligned} Y_{+6} &= 1.165 - 1.167\text{Log}(6.5) \\ &= 1.165 - 1.167(0.81291) \\ &= 0.21633 \end{aligned}$$

Converting this value to percent gives:

$$P_{+6} = 0.5856 \quad (\text{or } 58.56\%)$$

In addition, since the trainees had already been on board three years at the time of the program, their retention probability from 3 to 6 years is given by:

$$P = \frac{P_{+6}}{P_{+3}}$$

$$= \frac{0.5856}{0.7020} = 0.8342$$

Thus, the number of employees remaining from the original group of 100 trainees is:

$$\begin{aligned} N_{+6} &= P \times 100 \\ &= 0.8342 \times 100 = 83 \end{aligned}$$

To translate this number into an actual cost figure, use the formula:

$$AC = \frac{TC}{P}$$

Thus,

$$AC = \frac{\$1000}{.8342} = \$1199/\text{employee trained}$$

Thus, the three-year return on the training investment when employees who have already served three years are trained is $1199 per employee trained and still on board.

Conclusion - The return on a training investment can vary significantly based on the selection of employees to be trained. Management should be made aware of the cost differences involved so they can make an informed selection decision .

Question: 6 What are the group retention probabilities for employees in each length of service category?

Suppose that the log-probability equation for a given occupation is:

$$Y = 0.9534 - 0.9102\text{Log}(t)$$

To answer the question we need to calculate group retention probabilities; i.e., the probabilities that employees in given length of service categories during a given base year will still be on board at the end of the next year.

First consider those employees in their first year of service -- i.e., from 1 day to 1 year --during the base year. It is a simple matter to determine their group retention probability: just calculate the Y-value associated with the t-value of 1.5 years. (This is the same as calculating the retention from an original group one year later.) In this case we have:

$$\begin{aligned} Y &= 0.9534 - 0.9102\text{Log}(1.5) \\ &= 0.9534 - 0.9102(0.17609) \\ &= 0.79312 \end{aligned}$$

This Y-value corresponds to a retention probability of 0.7861.

To determine the probabilities for employees with more than one year of service, the first step is to iterate the log-probability equation for the midpoints of each length of service category beginning with 1 - 2 years; i.e., more than one but less than or equal to 2 years. (For convenience sake we will stop at the category 9 - 10 years. However, the method used would work for every length of service category.)

Iteration gives the following values:

L.O.S. (Years)	Midpoint	Retention Percent
1-2	1.5	78.61%
2-3	2.5	72.28
3-4	3.5	67.66
4-5	4.5	64.01
5-6	5.5	61.01
6-7	6.5	58.45
7-8	7.5	56.23
8-9	8.5	54.28
9-10	9.5	52.53

In addition, to get a retention probability for our last group, we also need to iterate the log-probability equation for the category 10 -11 years. This is:

10-11	10.5	50.95

The second, and final, step is to divide each year's retention percent into the succeeding year's percent. For example, the retention probability for employees with 1-2 years of service is:

$$\frac{72.28\%}{78.61\%} = 0.9195$$

Similar calculations for each length of service category give the following group retention probabilities:

L.O.S. (Years)	Retention Probability
0-1	0.7861
1-2	0.9195
2-3	0.9361
3-4	0.9461
4-5	0.9531
5-6	0.9580
6-7	0.9620
7-8	0.9653
8-9	0.9678
9-10	0.9699

These probabilities can be used to estimate the number of employees retained in each length of service category.

Question: 7 Using group retention probabilities, how many employees will be retained from a given group one year later?

Suppose that the length-of-service composition of a group of employees in a given occupation at the end of a given (or base) year is:

L.O.S. (years)	Number of Employees
0-1	100
1-2	81
2-3	76
3-4	71
4-5	67
5-6	64
6-7	61
7-8	58
8-9	55
9-10	52

This means that there are 100 employees in their first year of service, 81 in their second, etc. This is a total of 685 employees.

Suppose also that the log-probability equation for this group is:

$$Y = 0.9534 - 0.9102\text{Log}(t)$$

Using previously-explained techniques, the group retention probabilities associated with this log-probability equation are:

L.O.S. (years)	Retention Probability
0-1	0.7861
1-2	0.9195
2-3	0.9361
3-4	0.9461
4-5	0.9531
5-6	0.9580
6-7	0.9620
7-8	0.9653
8-9	0.9678
9-10	0.9699

To estimate the number from each length-of-service category to be retained at the end of the next year, multiply the number in each category by the retention probability for that same category. For example, of the 81 employees in the 1-2 year category, an estimated 74 will be retained at the end of the next year (since 81 x 0.9195 = 74).

The following table displays this entire process. It shows (a) the number of employees in each length-of-service category at the end of the base year, (b) the retention probability for each category and (c) the resulting estimated number retained at the end of the next year. The corresponding figures in this sequence are linked by arrows and the estimates of employees retained are moved down to their new length-of-service category.

L.O.S. (years)	Base Year	Retention Prob.	Base Yr+1
0-1	100	(.7861)	
1-2	81	(.9195)	79
2-3	76	(.9361)	74
3-4	71	(.9461)	71
4-5	67	(.9531)	67
5-6	64	(.9580)	64
6-7	61	(.9620)	61
7-8	58	(.9653)	59
8-9	55	(.9678)	56
9-10	52	(.9699)	53
10-11			50

The estimated number of employees retained at the end of base year +1 is 634. The retention rate for the entire group is:

634/685 = .9255 (or 92.55%)

Conversely, the estimated number lost during base year +1 is 51 (685 - 634) and the loss rate is:

51/685 = .0745 (or 7.45%)

Question 8: How do changes in hiring levels affect a group's turnover rate?

Suppose that the log-probability equation for a given occupation is:

$$Y = 0.9534 - 0.9102 \mathrm{Log}(t)$$

Suppose further that over a period of years there has been a steady number of hires into the occupation, say 100 hires per year. Then the length of service distribution at the end of the base year for those hired over the past five years would probably look something like this:

Years Since Hire	Number Remaining
5	63
4	66
3	69
2	74
1	81

This gives a total of 353 employees.

To estimate the number of these employees who will still be on board at the end of next year (base year + 1), we can use the group retention probabilities associated with the log-probability equation. These are (as calculated in a previous question):

Years Since Hire	Retention Probability
5	0.9580
4	0.9531
3	0.9461
2	0.9361
1	0.9195

After multiplying the number in each length of service category by the corresponding retention probability, we have:

Years Since Hire	Number at End of Base Year	Number at End of Base Yr +1	Number Lost
5	63	60	3
4	66	63	3
3	69	65	4
2	74	69	5
1	81	74	7
	353	331	22

Thus, of the 353 starting employees, 22 will be lost during the year.

Now let's hold this figure constant and see what effect different levels of hiring will have on the occupation's turnover rate.

Suppose that during the base year the same pattern of hiring continues -- i.e., 100 employees are hired. Of these, 79 will be left by the end of base year + 1 (using the log-probability equation). This means that 21 of the new hires will be lost by the end of base year + 1. Combining this with the losses from the other length of service categories we have:

$$\begin{gathered} 21 + 22 = 43 \text{ losses} \\ \text{out of} \\ 100 + 353 = 453 \text{ employees} \end{gathered}$$

This gives a turnover rate of:

$$\frac{43}{453} = 0.0949 \quad \text{(or 9.49\%)}$$

Now suppose that instead of hiring 100 employees during the base year, management decides to hire 200. The number remaining from this group at the end of base year + 1 would be 157. Thus, 43 employees would be lost. Combining this figure with the losses from the other length of service categories (which remain the same) we have:

$$\begin{gathered} 43 + 22 = 65 \text{ losses} \\ \text{out of} \\ 200 + 353 = 553 \text{ employees} \end{gathered}$$

In this case the turnover rate would be:

$$\frac{65}{553} = 0.1175 \quad \text{(or 11.75\%)}$$

Continuing in the same manner, we get the following turnover rates for different levels of hiring:

Number Hired Base Year	Turnover Rate Base Year + 1
0	6.23%
50	8.19
100	9.49
200	11.75
300	13.17

Thus, increasing the number of new hires increases the turnover rate. Conversely, decreasing hires lowers the turnover rate. This is a logical result since most turnover occurs during the early years of service.

It is also worthwhile to note here that a freeze on hiring will lower a group's turnover rate. This is important to realize since decisions to freeze hiring are usually coupled with plans to let natural attrition take care of lowering employment levels. However, most calculations of the rate of this lowering of employment are based on previous attrition rates and, thus, are usually wrong. The lowering of attrition rates in these cases must be taken into account.

Similarly, in a RIF situation, turnover rates are going to decrease both because no new hiring is done and because the first employees to be RIFed are those with fewer years of service.

APPENDIX C

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LOGPRO

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LOGPRO

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

LOGPRO TECHNICAL ANALYSIS

Introduction

The computer terminal program LOGPRO was developed to enable users to:

(1) perform a quantitative analysis of group turnover and retention rates (based on employee length of service); and

(2) use the results of this analysis to project these rates into the future.

Since the majority of all vacancies are caused by turnover (rather than by growth), the accuracy of projections of future hiring needs is directly related to the accuracy of turnover/retention analysis and projection.

LOGPRO and its two subprograms (ANDXP and ANDPX) together form a comprehensive, self-contained unit which completes all the necessary statistics required for the analysis and projection of turnover/retention rates. This means that a user of this set of programs need not have an extensive statistical background in order to successfully utilize and evaluate its results.

Data Collection and Input

For use in LOGPRO, turnover/retention data are obtained by determining the numbers or percent of employees retained from a given group of hires after specified lengths of service. More precisely, data for the LOGPRO analysis method are collected by cohort group; i.e., by following the retention behavior of the same group of employees over time. This type of data is also known as longitudinal data.

Ideally, cohort groups should be composed of employees in the same occupation. However, if the number of employees in one or more occupations is small, then it is possible to combine like occupations (e.g., all scientists). In addition, there are techniques available which will statistically determine which occupations may be grouped together. (See documentation for LPTEST.)

From each cohort to be analyzed, two types of information must be recorded:

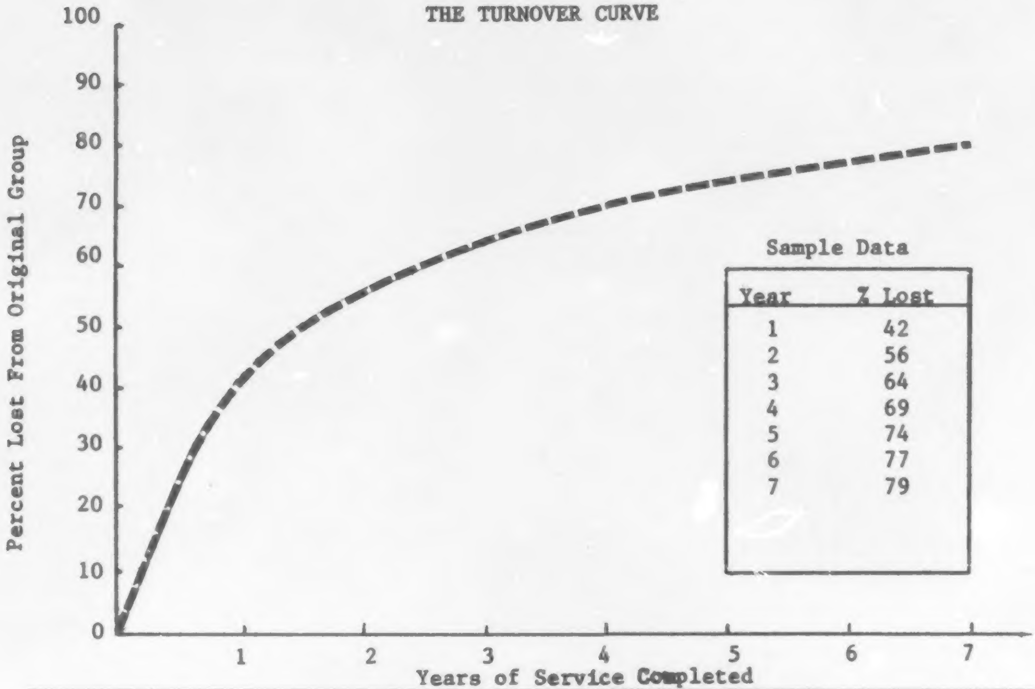
(1) The number of employees in the original group of hires; and

(2) The number (or percent) of these employees retained after given lengths of time.

The "original group" is composed of employees hired during a specific span of time. This span should be limited to one year or less. The "given lengths of time" could be any standard time units (years, months, days) after the end of the original time span. Note that, since

FIGURE 1

THE TURNOVER CURVE

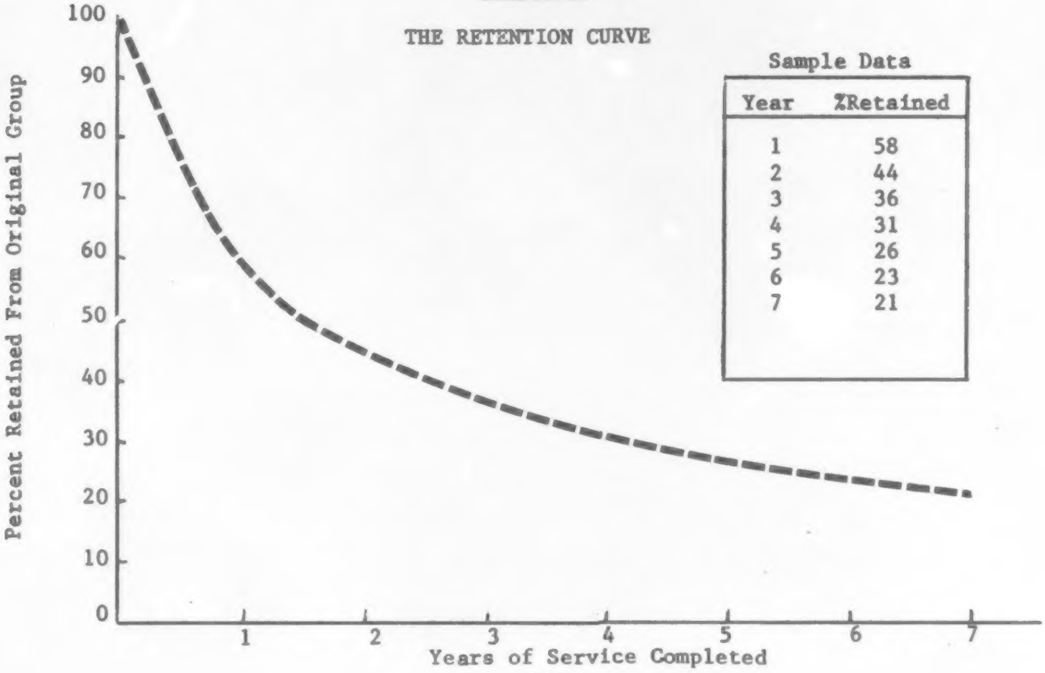


Sample Data

Year	% Lost
1	42
2	56
3	64
4	69
5	74
6	77
7	79

FIGURE 2

THE RETENTION CURVE



Sample Data

Year	%Retained
1	58
2	44
3	36
4	31
5	26
6	23
7	21

the original group will contain employees with different lengths of service (from 1 day to 1 year), there are several options involved in assigning values to the time units.

For example, consider a group of employees hired during a given fiscal year ending September 30th. And suppose that retention data for this group are recorded on succeeding September 30ths. The time values for these data points may be:

(a) 1 year, 2 years, etc. (calculated from September 30th to September 30th); or

(b) 1.5 years, 2.5 years, etc. (assuming an even distribution of hires over a year and using the midpoint; i.e., 0.5); or

(c) 1.X years, 2.X years, etc. (where X is a factor based upon specific hiring patterns, such as more hiring done at the end of a fiscal year); or

(d) 1.Y years, 2.Y years, etc. (where Y represents the actual average length of service of the original group at the end of the first fiscal year).

The Shape of the Curve

The key characteristic of a turnover curve can be summed up in a simple rule of thumb:

> Between two-thirds and three-fourths of all of the turnover losses that will ever take place from any given group of hires will have taken place by the end of the third year after hiring. And of that total, roughly half will have taken place by the end of the first year alone.

What this says is that:

(a) Turnover is a function of length of service; and

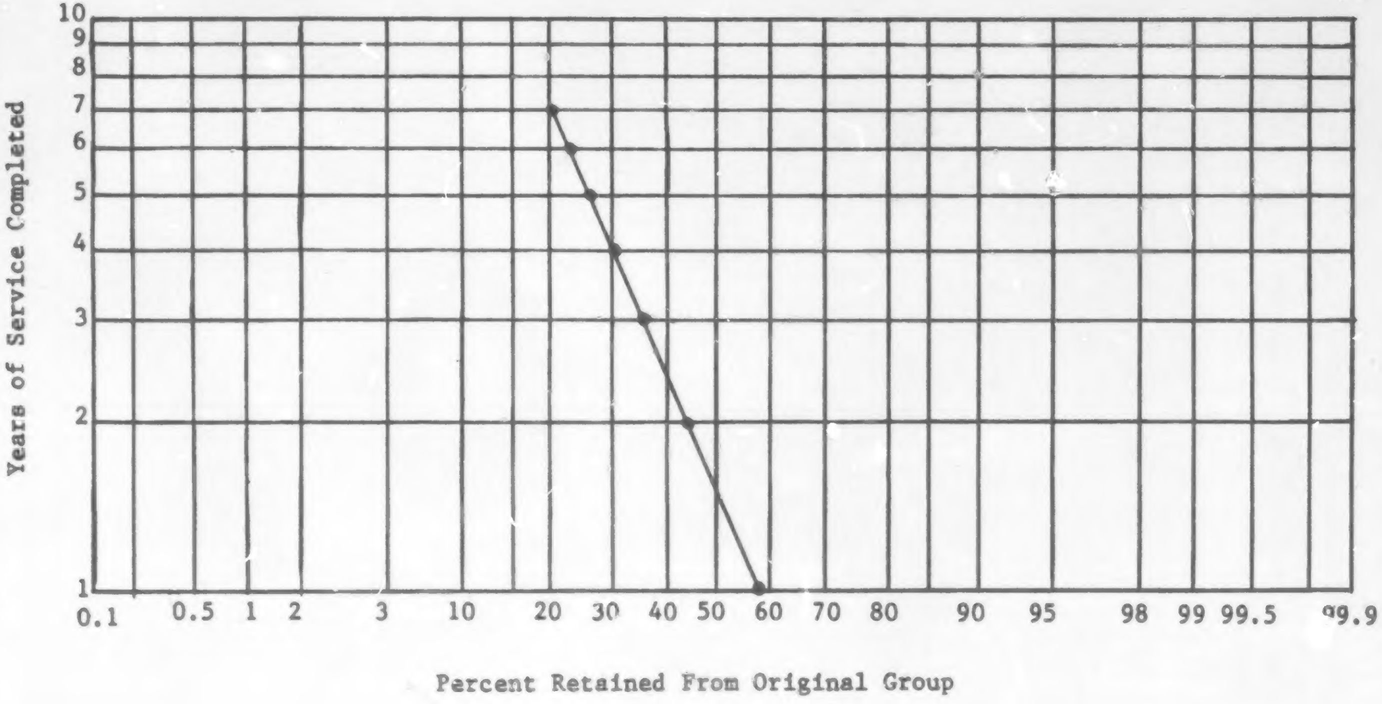
(b) Most turnover occurs during the early years of employment.

This key characteristic of the turnover curve gives it a particularly characteristic shape. In Figure 1, "years of service completed" is plotted along the horizontal axis and "percent lost from original group" is plotted along the vertical axis. It is immediately observable that the curve bends sharply during the first three years and then levels off.

In the same manner, a graph of a retention curve (since retention is simply "number of employees in the original group" minus "employees lost from the group") shows the same characteristics. Figure 2 plots the retention curve associated with the turnover curve in Figure 1. In

FIGURE 3

THE LOG-PROBABILITY LINE



this case, the vertical axis represents "percent of the original group retained." For purposes of the following analysis discussion, the retention curve will be used.

The simplest way to analyze such a curve is to transform the relevant variables so that the data points form a straight line. Then analysis can be done by the use of a least squares technique. In the case of the retention curve, the transformations are:

(a) *Percent of employees retained* is transformed into *number of standard deviations from the mean of a normal curve*; and

(b) *Years of service* is transformed into the *logarithm of years* of service.

These two transformations give this analysis method its name: the *log-probability* technique.

When the values in Figure 2 are plotted on log-probability graph paper, the result is a straight line (Figure 3).

Basic Assumptions

The basic assumptions behind the use of the log-probability technique for retention analysis are threefold.

First, the given group possesses a normal distribution of factors motivating group retention and departure behavior. This assumption is based upon the extreme complexity of the interaction between an employee and his or her job situation. Any one person can be motivated by any number of "stay" and "go" factors. For analysis purposes, the significant aspect of this multiplicity of retention-producing factors is not their "identity, but their *distribution* among the workgroup members. New hires with extremely "anti-this-job" attitudes, for example, could be expected to leave immediately, *regardless* of what the particular reasons for their individual attitudes might be. Those who were very "pro-this-job" would stay, regardless of *their* reasons. And the rest would be found somewhere between. Under this schema, then, the very multiplicity (and the mutual independence) of the employee retention/loss factors which so impossibly complicates their identification should ensure that a frequency distribution of their appearance in the workgroup should approach that of the normal curve." 1/

Second, the behavior resulting from these retention motivating factors will be displayed logarithmically, rather than arithmetically, over time. This assumption is based upon the distribution of turnover as expressed in the previously mentioned rule of thumb. The largest bulk of turnover

---

1/ Clark, H.L., "Problems and Progress in Civil Service Manpower Planning in the United States." In: *Manpower Planning Models* (Clough, Lewis & Oliver Eds), London, English Universities Press, 1974, pp. 227-239.

losses occurs in the first years of service with successively fewer losses occurring as years go by. This is a logarithmic progression.

Third, the situational factors affecting the retention behavior during the initial period of empirical observation will continue substantially unchanged throughout the remaining period of the projection. This assumption requires that there be no extreme change in the working situation of the cohort groups. Such changes would include large-scale reorganizations or program changes.

Statistical Analysis

The exact mathematical form of the retention curve (also known as a "decay" curve) is given by the linear equation:

$$Z(Y) = A + B \operatorname{Log}X \qquad (1)$$

where:

X = length of service completed.

Y = percent of original group retained at point X.

$Z(Y)$ = location on the normal curve of the retention point Y, expressed in standard deviations distance from the mean.

In order to obtain the data which can be used to perform the linear regression which will determine the values of A and B in equation (1), LOGPRO makes the aforementioned transformations of the inputted retention data (Page 5).

The first transformation uses the accumulative normal distribution function to determine both (a) the standard deviation value associated with a given retention percentage (used in least squares calculation) and (b) the retention percentage value associated with a given standard deviation (used in curve iteration and goodness of fit calculations). The mathematical formulas which accomplish this basic two-way transformation are done by LOGPRO's two subprograms ANDXP and ANDPX.

ANDXP - This subprogram uses an approximation of the accumulative normal distribution function to estimate P(percent retained) from a given value of X(standard deviations from the mean).

The form of this approximation 2/, using $X_1 = |X|$, is:

$$P(X_1) = 1 - Z(X_1)(b_1t + b_2t^2 + b_3t^3 + b_4t^4 + b_5t^5) + e(X_1) \qquad (2)$$

2/ Abramowitz, M. and Stegun I.A., Eds, Handbook of Mathematical Functions, AMS55, 9th, National Bureau of Standards, 26.2.17, p. 932.

where:

$$Z(x_1) = \frac{1}{\sqrt{2\pi}} e^{\left(\frac{-X_1^2}{2}\right)}$$

$$t = \frac{1}{1 + .2316419X_1}$$

$b_1 = 0.3193815$

$b_2 = -0.3565638$

$b_3 = 1.7814780$

$b_4 = -1.8212560$

$b_5 = 1.3302744$

If the value of the inputted X is greater than or equal to zero, then the value $P(X_1)$ calculated from equation (2) is returned to the main program. If X is less than zero, the the quantity $(1 - P(X_1))$ is returned.

The error function ($e(X_1)$) is accurate to ± 7.5×10^{-8}.

ANDPX - This subprogram uses an approximation of the inverse accumulative normal distribution function to estimate X from a given value of P. The form of this approximation 3/ is:

$$X(P) = t - \left(\frac{a_0 + a_1t + a_2t^2}{1 + b_1t + b_2t^2 + b_3t^3}\right) + e(P) \qquad (3)$$

where:

$$t = \sqrt{\ln \frac{1}{P^2}}$$

$a_0 = 2.515517$ $\qquad$ $b_1 = 1.432788$

$a_1 = 0.802853$ $\qquad$ $b_2 = 0.189269$

$a_2 = 0.010328$ $\qquad$ $b_3 = 0.001308$

3/ Ibid., 26.2.23, p. 933.

If the inputted value of P is greater than or equal to 0.5 (i.e, 50%), then the actual value of P is used in the calculation of t. If P is less than 0.5, then the quantity (1 - P) is used.

The error function (e(P)) for this approximation is accurate to ± 4.5×10^{-4}. To increase its accuracy for use by LOGPRO, ANDPX contains an additional refinement which makes use of the additional accuracy available in ANDXP.

To accomplish this refinement, the subprogram ANDPX:

(1) Takes the value of X computed using equation (3) and adds to it ± 0.0005 to determine two new values: X_1 and X_2. This establishes one interpolation range.

(2) Uses ANDXP to determine the percentage values associated with X_1 and X_2. This establishes a second interpolation range.

(3) Uses these two ranges to interpolate a new X value, X_3, to which it applies ANDXP and determines a percentage value, P_3.

(4) Computes the difference, D, between P and P_3. This difference is then tested to determine if it falls below 5×10^{-7}. If it does not, then it is recomputed (using new X_i and P_i values) until it does.

(5) Returns the X_i value associated with the minimum difference to LOGPRO.

This series of steps increases the accuracy of ANDPX to seven places.

Computations

After transforming the inputted retention data into least squares form, LOGPRO performs a series of computations:

(1) The transformed data are used to calculate a log-probability retention equation.

(2) The value of the regression equation is calculated for each retention point. This value is changed into both estimated number and percent retained.

(3) The squared deviation of each retention point from the line of regression is calculated and used to compute the standard deviation of the fit.

(4) At operator option, the log-probability regression equation can be projected into the future. Standard extrapolation procedures are used. The operator specifies the desired projection parameters.

(5) If the number of retention points inputted is greater than 2, the 95% confidence range is calculated for each projected point (using the standard deviation of the fit). The range values are computed in both number and percent form.

Applications to Other Areas

Log-probability retention lines provide insight into a number of management and personnel problems. For example, they can be used to estimate:

- Turnover expected in specific planning periods (e.g., 17% in the first year, 9% in the second year, etc.).
- Current Hiring Needed in order to have specific numbers of workers on board on a particular date in the future (e.g., 160 hires now to have 100 on board five years hence).
- Budget necessary for current hires still present at each intermediate step toward target (e.g., 133 at start of second year, 120 at start of third year, etc.).
- Turnover Implications of selection method elected by management (e.g., first-year quit rate of 17% for a new hire, 6% for selection of 3-years-service employee, etc.).
- Training Cost implications of employee selection methods such as "three-year return on training investment" (comparative training investment per worker on board three years after training is appreciably higher for a new outside hire than for an on-board employee reassigned or promoted for training).
- Personnel Cost implications of budgetary constraints such as hiring freezes, employment cutbacks, etc. (Usually result in a decrease in the number of low length-of-service employees in the workforce, thus decreasing employee turnover rates).

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LOGPRO

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

"LOGPRO" IS A FORTRAN IV PROGRAM WHICH PERFORMS LOG-PROBABILITY RETENTION ANALYSIS. IT REQUIRES TWO SUBPROGRAMS: ANDXP AND ANDPX.

DATA REQUIREMENTS:

THE "LOGPRO" PROGRAMMING SEQUENCE IS DESIGNED TO USE LONGITUDINAL "RETENTION OVER TIME" DATA. SUCH DATA ARE OBTAINED BY OBSERVING THE RETENTION BEHAVIOR OF A GROUP OF EMPLOYEES HIRED DURING A GIVEN PERIOD OF TIME, SAY, ONE YEAR. THESE SAME EMPLOYEES ARE THEN FOLLOWED OVER TIME TO DETERMINE HOW MANY (OR WHAT PERCENT) OF THEM REMAIN IN THEIR JOBS AFTER GIVEN PERIODS OF TIME; E.G., ONE YEAR, TWO YEARS, ETC. ANALYSIS OF THESE TYPES OF DATA IS ALSO KNOWN AS "COHORT ANALYSIS".

ALTHOUGH ANY "GROUP OF EMPLOYEES" MAY BE USED IN THIS PROGRAM, COHORT ANALYSIS OF THIS TYPE IS MORE ACCURATE IF THE EMPLOYEES IN THE GROUP(S) SELECTED ARE IN THE SAME OR SIMILAR OCCUPATIONS. FOR EXAMPLE, ONE GROUP COULD BE ALL CLERK-TYPISTS HIRED IN FY73. ANOTHER COULD BE ALL SCIENTISTS HIRED IN FY73.

THE PERIODS OF TIME USED COULD BE ANY STANDARD TIME UNITS (YEARS, MONTHS, DAYS) OR ANY FRACTION THEREOF. BECAUSE MOST COHORT GROUPS WILL CONSIST OF EMPLOYEES WITH VARYING LENGTHS OF SERVICE (FROM 1 DAY TO 1 YEAR), AN AVERAGING FACTOR CAN BE ADDED TO THESE TIME UNITS TO MORE CLOSELY APPROXIMATE THE ACTUAL AVERAGE LENGTH OF SERVICE OF THE COHORT (E.G., FOR A GROUP OF EMPLOYEES HIRED IN AN EVEN DISTRIBUTION DURING A FISCAL YEAR, THE FACTOR WOULD BE 0.5 YEARS).

THE CURVE FORM USED IN "LOGPRO" IS KNOWN AS THE "LOG-PROBABILITY" CURVE. THIS CURVE IS USED BECAUSE IT MOST CLOSELY APPROXIMATES ACTUAL RETENTION BEHAVIOR.

HYPOTHETICAL DATA SET:

THE FOLLOWING IS A HYPOTHETICAL SET OF DATA OF THE TYPE REQUIRED FOR THE "LOGPRO" PROGRAMMING SEQUENCE:

> SUPPOSE 110 CLERK-TYPISTS WERE HIRED DURING FY1972 AND SUPPOSE THESE HIRES WERE MADE EVENLY THROUGHOUT THE FISCAL YEAR AND SUPPOSE THAT

AT THE END OF FY:	THE NUMBER RETAINED WAS:
1973	58 (OR 52.73%)
1974	44 (OR 40.00%)
1975	36 (OR 32.73%)

THE AVERAGING FACTOR FOR THIS DATA SET IS 0.5 YEARS. THIS SET OF DATA WILL BE REFERRED TO THROUGHOUT THIS MANUAL.

EXECUTION COMMANDS:

TO BEGIN EXECUTION OF THE "LOGPRO" PROGRAMMING SEQUENCE, AN OPERATOR WILL PERFORM A CHAIN OF EXECUTION COMMANDS. THE ACTUAL FORM OF THESE COMMANDS WILL DEPEND ON THE TIME-SHARING

* = COMPUTER WRITTEN

SYSTEM BEING USED. IN GENERAL, THESE COMMANDS WILL PERFORM THE FOLLOWING OPERATIONS:

- CALL UP "LOGPRO" AND ITS TWO SUBPROGRAMS "ANDXP" AND "ANDPX" AND TRANSLATE THEM INTO MACHINE LANGUAGE. THIS IS THE COMPILATION PHASE.
- LOAD THE COMPILED PROGRAM AND SUBPROGRAMS INTO THE CENTRAL PROCESSING AREA AND START PROGRAM RUN. THIS IS THE EXECUTION PHASE.

DATA ENTRY:

(NOTE THAT ALL OPERATOR-ENTERED RESPONSES TO COMPUTER-WRITTEN COMMANDS ARE FOLLOWED BY A CARRIAGE RETURN.)

THE RUN OF THE "LOGPRO" ANALYSIS SEQUENCE BEGINS IN THE FOLLOWING MANNER:

```
* PLEASE ENTER THE NUMBER OF YEARS FOR WHICH RETENTION DATA ARE
* AVAILABLE
```

THE OPERATOR RESPONDS TO THIS COMMAND BY ENTERING THE NUMBER OF YEARS (OR, IF DATA ARE IN OTHER THAN YEARS, THE NUMBER OF TIME PERIODS) FOR WHICH COHORT-TYPE LONGITUDINAL DATA ARE AVAILABLE. IN THE CASE OF THE HYPOTHETICAL DATA SET, THIS ANSWER WOULD BE "3".

NEXT THE COMPUTER ASKS:

```
* PLEASE ENTER THE NUMBER IN THE STARTING GROUP
```

AT THIS POINT, THE OPERATOR ENTERS THE NUMBER OF EMPLOYEES WHO COMPOSED THE ORIGINAL GROUP OF HIRES FOR WHICH COHORT ANALYSIS IS BEING PERFORMED. USING THE HYPOTHETICAL DATA SET WHERE 110 EMPLOYEES WERE HIRED IN FY 1972 AND FOLLOWED DURING THE NEXT 3 YEARS, THE RESPONSE TO THE ABOVE QUESTION WOULD BE "110".

NEXT:

```
* WILL INPUT DATA BE IN
*    (1) NUMBER OR
*    (2) PERCENT
*    (ANS 1 OR 2)
```

IF THE RETENTION DATA ARE IN THE FORM "NUMBER OF EMPLOYEES RETAINED FROM THE ORIGINAL GROUP", THEN THE RESPONSE TO THE ABOVE COMMAND IS "1".

ON THE OTHER HAND, IF THE DATA ARE IN THE FORM "PERCENT OF EMPLOYEES RETAINED FROM THE ORIGINAL GROUP", THEN THE RESPONSE IS "2".

NEXT THE COMPUTER ASKS FOR THE RETENTION DATA IN THE FOLLOWING MANNER:

* = COMPUTER WRITTEN

```
* PLEASE ENTER THE VALUES OF X (TIME) AND Y (NUMBER OR PERCENT
* RETAINED BY TIME X) IN THE FORM X,Y SHOWING ALL DECIMAL POINTS
```

AT THIS POINT, THE OPERATOR ENTERS THE RETENTION-OVER-TIME DATA WHICH ARE TO BE ANALYZED. THESE DATA ARE ENTERED AS (X,Y) PAIRS WHERE X REPRESENTS A UNIT OF TIME AND Y REPRESENTS THE NUMBER OR PERCENT OF THE ORIGINAL GROUP OF EMPLOYEES REMAINING AT TIME X.

IF THE X-VALUES ARE IN YEARS,THEN THE ENTRIES FOR THESE VALUES ARE IN THE FORM 1.X, 2.X, 3.X, ETC. (WHERE X REPRESENTS THE AVERAGING FACTOR). IF MONTHS ARE USED, THEN THE ENTRIES CAN BE IN THE FORM (A) 3.+Y, 6.+Y, 9.+Y MONTHS (WHERE Y = X TIMES 12) OR (B) .25+X, .5+X, .75+X YEARS. THE X-VALUES DO NOT HAVE TO BE EQUALLY SPACED.

IF THE ANSWER TO THE PREVIOUS QUESTION HAD BEEN "1", THEN THE Y VALUES ARE IN THE FORM "NUMBER OF EMPLOYEES RETAINED". IN THIS CASE THE ACTUAL NUMBERS RETAINED ARE ENTERED INTO THE PROGRAM (SHOWING ALL DECIMAL POINTS). USING THE HYPOTHETICAL DATA SET, THE X,Y ENTRIES WOULD BE:

```
1.5,58.
2.5,44.
3.5,36.
```

IF THE ANSWER TO THE PREVIOUS QUESTION HAD BEEN "2", THEN THE Y VALUES ARE IN THE FORM "PERCENT OF EMPLOYEES RETAINED". IN THIS CASE, THE PERCENTAGES ARE ENTERED IN THEIR DECIMAL FORM; E.G., 52.73% IS ENTERED AS .5273. USING THE HYPOTHETICAL DATA SET, THE X,Y ENTRIES WOULD BE:

```
1.5,.5273
2.5,.4000
3.5,.3273
```

LOG-PROBABILITY ANALYSIS:

THE COMPUTER THEN PERFORMS ALL THE NECESSARY LOG-PROBABILITY ANALYSIS FUNCTIONS. THESE INCLUDE:

- FITTING THE LOG-PROBABILITY CURVE;
- COMPUTING THE LOG-PROBABILITY CURVE VALUES (BOTH IN NUMBER AND PERCENT) FOR EACH TIME PERIOD ENTERED; AND
- COMPUTING THE STANDARD DEVIATION OF THE LOG-PROBABILITY FIT (IF THE NUMBER OF OBSERVATIONS IS GREATER THAN 2).

THE COMPUTER THEN PRINTS THE "TABLE OF LOG-PROBABILITY ANALYSIS RESULTS" WHICH SUPPLIES THE OPERATOR WITH ALL OF THE RESULTS OF

THE ANALYSIS. IN ADDITION TO PRINTING OUT THE ACTUAL AND CURVE VALUES, THE COMPUTER ALSO PRINTS OUT THE A AND B VALUES OF THE LOG-PROBABILITY EQUATION AND THE VALUE OF THE STANDARD DEVIATION OF THE FIT. NOTE THAT THESE LATTER 3 VALUES ARE IN THE FORM "STANDARD DEVIATIONS FROM THE MEAN OF THE NORMAL CURVE". (SEE PAGE 7).

THE FORMAT OF THIS PRINTOUT (USING THE HYPOTHETICAL DATA SET) OF RESULTS IS:

```
*XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
*
*
*           TABLE OF LOG-PROBABILITY ANALYSIS RESULTS
*          ------------------------------------------
*
*                 --ACTUAL DATA--   --L - P CURVE--
*           YEAR   NUMBER  PERCENT   NUMBER  PERCENT
*
*
*           1.50      58.    52.73    57.88    52.62
*           2.50      44.    40.00    44.30    40.27
*           3.50      36.    32.73    35.83    32.57
*
*          ------------------------------------------
*
*         THE NUMBER IN THE STARTING GROUP WAS:   110
*
*         THE LOG-PROBABILITY EQUATION IS
*           Y =   0.31317 + ( -1.40587) X
*
*         THE STANDARD DEVIATION IF FIT IS 0.006205
*
*
*XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

PROJECTIONS:

THE NEXT PORTION OF THE "LOGPRO" PROGRAMMING SEQUENCE DEALS WITH THE PROJECTION OF THE PREVIOUSLY CALCULATED LOG-PROBABILITY CURVE. A PROJECTION IS A "THROWING FORWARD" OF THE TREND OF PAST DATA INTO THE FUTURE.

IN "LOGPRO", THE USE OF PROJECTION TECHNIQUES ALLOWS AN OPERATOR TO FORM ESTIMATES OF THE NUMBER OF EMPLOYEES FROM THE ORIGINAL GROUP WHO WILL BE RETAINED IN THEIR JOBS AT GIVEN POINTS IN THE FUTURE. SINCE SUCH PROJECTION FIGURES ARE "ESTIMATES" AND NOT EXACT VALUES, "LOGPRO" ALSO COMPUTES (FOR GROUPS WITH MORE THAN TWO TIME VALUES ENTERED) THE 95% CONFIDENCE RANGE VALUES IN BOTH NUMBER AND PERCENT. THIS MEANS THAT, FOR A GIVEN PROJECTION

POINT, THE PROGRAM WILL CALCULATE A RANGE OF VALUES WITHIN WHICH THE ACTUAL NUMBER RETAINED AT A GIVEN POINT WILL FALL 95% OF THE TIME.

THE PROJECTIONS SEQUENCE BEGINS AS FOLLOWS:

```
* IS A PROJECTION DESIRED? (YES=Y, NO=N)
```

IF NO SUCH PROJECTION IS WANTED, THE RESPONSE TO THE QUESTION IS "N". THE COMPUTER WILL THEN PROCEED TO THE NEXT PORTION OF THE PROGRAM.

IF A PROJECTION IS DESIRED, THE RESPONSE IS "Y". THE COMPUTER THEN PRINTS OUT THE FOLLOWING INSTRUCTIONS:

```
* PLEASE ENTER THE MINIMUM AND MAXIMUM VALUES OF X DESIRED
* PLUS THE DESIRED X-INCREMENT (E.G., .5 YEAR, 1 YEAR, ETC.)
* IN THE FORM: MIN, MAX, INCREMENT (E.G., 5.,15.,1.).
* PLEASE SHOW ALL DECIMAL POINTS.
```

AT THIS POINT THE OPERATOR ENTERS THREE VALUES:

1. THE FIRST (I.E., MINIMUM) TIME VALUE FOR WHICH A PROJECTION IS WANTED; E.G., 5TH YEAR;

2. THE LAST (I.E., MAXIMUM) TIME VALUE FOR WHICH A PROJECTION IS WANTED; E.G., 10TH YEAR; AND

3. THE LENGTH OF THE TIME INTERVAL BETWEEN SUCCESSIVE PROJECTION POINTS (I.E., INCREMENT); E.G.,1 YEAR.

TO ALLOW FOR THE USE OF FRACTIONAL TIME PERIODS (SUCH AS THE USE OF AN AVERAGING FACTOR), EACH PIECE OF DATA ENTERED HERE MUST INCLUDE A DECIMAL POINT.

FOR EXAMPLE, THE VALUES LISTED ABOVE WOULD BE ENTERED AS:

```
5.,10.,1.
```

THE COMPUTER THEN PERFORMS THE CALCULATIONS NECESSARY TO PROJECT THE LOG-PROBABILITY CURVE PREVIOUSLY CALCULATED.

IN ADDITION, IF THE NUMBER OF RETENTION DATA SETS ENTERED IS GREATER THAN 2, THE PROGRAM CALCULATES THE 95% CONFIDENCE RANGE VALUES FOR EACH PROJECTED POINT. THESE ARE CALCULATED FOR BOTH NUMBER AND PERCENT VALUES. THUS, FOR EACH PROJECTION POINT, THE PROGRAM CALCULATES AN "EXPECTED VALUE" (OR "PROJECTED ESTIMATE") AND A 95% RANGE OF VALUES (LOWEST AND HIGHEST).

THE PRINTED OUTPUT FOR THIS PORTION OF THE PROGRAM IS (USING THE HYPOTHETICAL DATA SET AND THE PROJECTION LIMITS 5.5,10.5,1.):

* = COMPUTER WRITTEN

```
* ----------------TABLE OF PROJECTED VALUES-----------------
*
*  YEAR ----NUMBER PROJECTED---- -----PERCENT PROJECTED-----
*       EXPECTED                  EXPECTED
*        VALUE    ---95% RANGE---  VALUE     ----95% RANGE----
*
*  5.50      26.      25. -  26.     23.34      22.96 - 23.72
*  6.50      22.      22. -  23.     20.34      19.99 - 20.69
*  7.50      20.      19. -  20.     17.96      17.63 - 18.28
*  8.50      18.      17. -  18.     16.02      15.73 - 16.33
*  9.50      16.      16. -  16.     14.43      14.15 - 14.71
* 10.50      14.      14. -  15.     13.08      12.82 - 13.35
*
* -----------------------------------------------------------
```

IF THE NUMBER OF RETENTION DATA SETS IS LESS THAN OR EQUAL TO 2, THEN THE 95% RANGE CANNOT BE CALCULATED AND ONLY EXPECTED VALUES ARE CALCULATED AND PRINTED OUT.

IN THIS CASE, THE OUTPUT WOULD BE (USING ONLY THE FIRST TWO ELEMENTS OF THE HYPOTHETICAL DATA SET AND THE SAMPLE PROJECTION LIMITS ABOVE):

```
* ----------------TABLE OF PROJECTED VALUES-----------------
*
*  YEAR ----NUMBER PROJECTED---- -----PERCENT PROJECTED-----
*  5.50             25.                   22.66
*  6.50             22.                   19.62
*  7.50             19.                   17.22
*  8.50             17.                   15.28
*  9.50             15.                   13.69
* 10.50             14.                   12.35
*
* -----------------------------------------------------------
```

RECYCLING:

NEXT THE COMPUTER ASKS:

* RUN AGAIN WITH A DIFFERENT DATA SET? (Y OR N)

IF THE OPERATOR WISHES TO RUN ANOTHER SET OF DATA THROUGH THE "LOGPRO" PROGRAMMING SEQUENCE THEN THE ANSWER TO THIS QUESTION IS "Y" (I.E., YES).

IF ALL DATA SETS HAVE ALREADY BEEN ANALYZED THEN THE ANSWER IS "N" (I.E., NO).

* = COMPUTER WRITTEN

SUBPROGRAMS:

THE "LOGPRO" PROGRAMMING SEQUENCE USES TWO SUBPROGRAMS WHICH CONSTITUTE THE BASIC STATISTICAL TECHNIQUE NEEDED FOR LOG-PROBABILITY ANALYSIS.

THE FIRST, KNOWN AS "ANDPX", TRANSFORMS "PERCENT RETAINED" INTO "NUMBER OF STANDARD DEVIATIONS FROM THE MEAN OF THE NORMAL CURVE". THE SECOND, "ANDXP", IS THE INVERSE OF "ANDPX". (SEE "LOGPRO" TECHNICAL ANALYSIS FOR A DETAILED EXPLANATION OF THESE TRANSFORMATIONS.)

APPENDIX C-3 PROGRAM LISTING

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LOGPRO

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

```
C          -----LOG-PROBABILITY ANALYSIS OF GROUP RETENTION TREND-----
C
C          CALCULATES THE LOG-PROBABILITY EQUATION FROM LONGITUDINAL
C          RETENTION DATA WHICH IS EITHER IN THE FORM (A) "NUMBER OF
C          EMPLOYEES RETAINED" OR (B) "PERCENTAGE (OF ORIGINAL GROUP)
C          OF EMPLOYEES RETAINED". ALSO CALCULATES THE STANDARD DEVIA-
C          TION OF FIT.
C          IN ADDITION, IF DESIRED, PROJECTS AND/OR INTERPOLATES THE
C          LOG-PROBABILITY CURVE AND CALCULATES 95 % CONFIDENCE RANGE
C          VALUES FOR EACH PROJECTION POINT.
C          REQUIRES FUNCTIONS ANDPX AND ANDXP.
C
         LOGICAL INV
         DIMENSION X(10),Y(10),XL(10),YS(10),A(10),AY(10),AA(10),NX(36)
         1  ,NDH(36)
C        +,IV(10,50),IYR(10)
         DATA (NX(I),I=1,36)/36*´XX´/,(NDH(I),I=1,36)/36*´--´/
25       INV=.FALSE.
         WRITE (5,27)
27       FORMAT(//1X,´PLEASE ENTER THE NUMBER OF YEARS FOR ´
         1  ´WHICH RETENTION DATA ARE AVAILABLE´,/)
30       READ (5,40)N
         IF(N.GT.2)INV=.TRUE.
         WRITE (5,32)
32       FORMAT (//1X,´PLEASE ENTER THE NUMBER IN THE STARTING GROUP´,/)
         READ (5,34)NO
34       FORMAT (I6)
         AN = NO
40       FORMAT (I2)
         WRITE (5,42)
42       FORMAT (//,1X,´WILL INPUT DATA BE IN´,/,3X,´(1) NUMBER OR´,/,
         1  3X,´(2) PERCENT´,/,3X,´(ANS 1 OR 2)´,/)
         READ (5,44)NP
44       FORMAT (I2)
         WRITE (5,47)
47       FORMAT (//1X,´PLEASE ENTER THE VALUES OF X (TIME) AND Y (NUMBER´
         1  ´ OR PERCENT´)
         WRITE(5,51)
51       FORMAT(1X,´RETAINED BY TIME X) IN THE FORM X,Y SHOWING ALL DECI´
         1  ´MAL POINTS´,/)
         DO 100 I=1,N
         READ (5,*)X(I),Y(I)
70       FORMAT(2F10.4)
         IF(NP.EQ.1) Y(I)=Y(I)/AN
         XL(I) = ALOG10(X(I))
         YS(I) = Y(I)
         YS(I) = ANDPX(YS(I))
100      CONTINUE
         SUMX = 0.0
         SUMY = 0.0
         SUMXY = 0.0
         SUMX2 = 0.0
         SUMY2 = 0.0
         DO 220 I = 1,N
         SUMX = SUMX + XL(I)
```

```
      SUMY = SUMY + YS(I)
      SUMXY = SUMXY + XL(I)*YS(I)
      SUMX2 = SUMX2 + XL(I)*XL(I)
      SUMY2 = SUMY2 + YS(I)*YS(I)
220   CONTINUE
      EN = N
      S1 = EN*SUMX2-SUMX*SUMX
      S2 = EN*SUMXY- SUMX*SUMY
      SL = S2/S1
      YIN = SUMY/EN-(SL*SUMX)/EN
      SUMD=0.0
380   DO 450 I = 1,N
      A(I) = YIN + SL*XL(I)
      SUMD=SUMD+(A(I)-YS(I))**2
      A(I) = ANDXP (A(I))
450   CONTINUE
      ASD=0.0
      IF(.NOT.INV) GO TO 522
      ASD=SQRT(SUMD/(EN-1.))
522   DO 527 I = 1,N
      AY(I) = AN*Y(I)
      AA(I) = AN*A(I)
527   CONTINUE
528   WRITE(5,529)(NX(I),I=1,36)
529   FORMAT(/////,1X,36A2)
      WRITE(5,531)
531   FORMAT(//,17X,'TABLE OF LOG-PROBABILITY ANALYSIS RESULTS')
      WRITE(5,823)(NDH(I),I=1,21)
823   FORMAT(16X,21A2)
      WRITE(5,534)
534   FORMAT(/,24X,'--ACTUAL DATA--',3X,'--L - P CURVE--')
      WRITE(5,536)
536   FORMAT(17X,'YEAR',3X,'NUMBER  PERCENT',3X,'NUMBER  PERCENT'
      1  ,//)
      DO 543 I=1,N
      Y(I)=Y(I)*100.
      A(I)=A(I)*100.
      WRITE(5,542)X(I),AY(I),Y(I),AA(I),A(I)
542   FORMAT(17X,F4.2,3X,F6.0,2X,F7.2,3X,F6.2,2X,F7.2)
543   CONTINUE
      WRITE(5,932)(NDH(I),I=1,21)
932   FORMAT(/,16X,21A2)
      WRITE(5,934)NO
934   FORMAT(/,15X,'THE NUMBER IN THE STARTING GROUP WAS:',I6)
      WRITE(5,892)
892   FORMAT(/,15X,'THE LOG-PROBABILITY EQUATION IS')
      WRITE(5,894)YIN,SL
894   FORMAT(17X,'Y = ',F9.5,' + (',F9.5,') X')
      WRITE(5,896)ASD
896   FORMAT(/,15X,'THE STANDARD DEVIATION OF FIT IS',F9.6)
      WRITE(5,907)(NX(I),I=1,36)
907   FORMAT(//1X,36A2)
530   WRITE (5,540)
540   FORMAT (///1X,'IS A PROJECTION DESIRED? (YES=Y, NO=N)',/)
      READ (5,560)L
```

```
560     FORMAT (A2)
        IF (L.EQ.'N') GO TO 690
        WRITE (5,590)
590     FORMAT (/1X,'PLEASE ENTER THE MINIMUM AND MAXIMUM VALUES OF'
     1  ' X DESIRED')
        WRITE(5,1182)
1182    FORMAT(1X,'PLUS THE DESIRED X-INCREMENT (E.G., .5 YEAR, 1 '
     1  'YEAR, ETC.)')
        WRITE(5,1185)
1185    FORMAT(1X,'IN THE FORM: MIN, MAX, INCREMENT (E.G., 5.,15.,1.).')
        WRITE(5,1188)
1188    FORMAT(1X,'PLEASE SHOW ALL DECIMAL POINTS.',/)
        READ (5,*) XMIN,XMAX,XINC
610     FORMAT (3F8.4)
        WRITE(5,1196)(NDH(I),I=1,16)
1196    FORMAT(///1X,8A2,'TABLE OF PROJECTED VALUES',8A2,'-',/)
        WRITE(5,1201)
1201    FORMAT(1X,' YEAR ----NUMBER PROJECTED---- -----PERCENT '
     1  'PROJECTED-----')
        IF(INV)WRITE(5,1206)
1206    FORMAT(7X,'EXPECTED',17X,'EXPECTED')
        IF(INV)WRITE(5,1208)
1208    FORMAT(8X,'VALUE',3X,'---95% RANGE---',2X,'VALUE',4X,
     1  '----95% RANGE----',/)
        QR = XMIN
        ASD=ASD*2.
630     ANS = YIN + SL*ALOG10(QR)
        ANS1 = ANDXP(ANS)
        R1=ANS+ASD
        R2=ANS-ASD
        ANN=ANS1*AN
        R1=ANDXP(R1)
        R2=ANDXP(R2)
        AN1=R1*AN
        AN2=R2*AN
        R1=R1*100.
        R2=R2*100.
        ANS1=ANS1*100.
        IF(.NOT.INV)WRITE (5,660)QR,ANN,ANS1
660     FORMAT (1X,F5.2,9X,F6.0,20X,F7.2)
        IF(INV)WRITE(5,665)QR,ANN,AN2,AN1,ANS1,R2,R1
665     FORMAT(1X,F5.2,2X,F6.0,2X,F6.0,' -',F5.0,4X,F6.2,3X,F7.2,
     1  ' -',F6.2)
        QR = QR + XINC
        IF (QR.LE.XMAX) GO TO 630
        WRITE(5,1282)(NDH(I),I=1,28)
1282    FORMAT(/1X,28A2,'-')
        WRITE(5,907)(NX(I),I=1,36)
690     WRITE (5,700)
700     FORMAT(///1X,'RUN AGAIN WITH A DIFFERENT DATA SET? (Y OR N)',/)
        READ (5,720)K
720     FORMAT (A2)
        IF (K.EQ.'Y') GO TO 25
        STOP
        END
        FUNCTION ANDPX(P)
C       ACCUMULATIVE NORMAL DISTRIBUTION FUNCTION.
C       APPROXIMATES X (+ OR - .000004) FROM P.
        DATA A0,A1,A2,B1,B2,B3/2.515517,.802853,.010328,1.432788,
        1  .189269,.001308/
        IF(P.GE.1.) X1=4.
        IF(P.GE.1.) GO TO 380
        IF(P.GT.0.0005) GO TO 70
        X1=-4.
        GO TO 380
70      IF (P.LT.0.5) GO TO 110
        E = SQRT(ALOG (1./(1.-P)**2))
        X1 = E-((A2*E+A1)*E+A0)/(((B3*E+B2)*E+B1)*E+1.)
        GO TO 140
110     P = 1.-P
        E = SQRT(ALOG(1./(1.-P)**2))
        X1 = -1.*(E-((A2*E+A1)*E+A0)/(((B3*E+B2)*E+B1)*E+1.))
        P = 1.-P
140     AX1 = X1 +0.0005
        AX2 = X1 -0.0005
        AP1 = ANDXP(AX1)
        AP2 = ANDXP(AX2)
        AI = (P-AP2)/(AP1-AP2)
        BX1 = AX2+AI*(AX1-AX2)
        D1 = P-ANDXP(BX1)
        IF (D1.LT.0.000001) GO TO 290
        AX3 = BX1+1.1*D1
        AX4 = BX1-1.1*D1
        AP3 = ANDXP(AX3)
        AP4 = ANDXP(AX4)
        AII = (P-AP4)/(AP3-AP4)
        BX1 = AX4+AII*(AX3-AX4)
        D1 = P-ANDXP(BX1)
290     IF (D1.LE.0.0000005) GO TO 370
        IF (D1.GT.0.0000005) GO TO 340
        BX1 = BX1+0.0000002
        D1 = ANDXP(BX1)
        GO TO 290
340     BX1 = BX1-0.0000002
        D1 = ANDXP(BX1)
        GO TO 290
370     ANDPX = BX1
        GO TO 390
380     ANDPX=X1
390     RETURN
        END
```

```
      FUNCTION ANDXP(X)
C     ACCUMULATIVE NORMAL DISTRIBUTION FUNCTION. APPROXIMATES P
C     FROM X (TO + OR - .0000001). (NBS-55, P. 932)
      X1 = ABS(X)
      T = 1./(1.+.2316419*X1)
      ANDXP = 1.-.3989423*EXP(-(X1**2)/2.)*(.3193815*T-.3565638*T**2
     1  +1.781478*T**3-1.821256*T**4+1.330274*T**5)
      IF (X.LT.0.0) ANDXP = 1.-ANDXP
      RETURN
      END
```

APPENDIX D

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LPFILE

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LPFILE

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

LPFILE TECHNICAL ANALYSIS [1]

Introduction

The computer terminal program LPFILE was developed to enable users to perform a quantitative analysis of turnover/retention rates for ungrouped, non-longitudinal data.

LPFILE and its three subprograms (ANDXP, ANDPX and DATE) together form a comprehensive, self-contained unit which completes all the necessary statistics required for the analysis of turnover/retention rates. This means that a user of this set of programs need not have an extensive statistical background in order to successfully utilize its results.

LPFILE calculates a log-probability equation from previously stored data files. The statistical transformations and basic assumptions of the longitudinal log-probability analysis method also apply to LPFILE. (See LOGPRO Technical Analysis, pages 3-8).

Data Collection and Input

LPFILE is designed to perform log-probability analysis when longitudinal data are not available. It does so by using computer files which contain certain basic data items on each employee in a given group. By utilizing data on every employee, LPFILE thus utilizes every available contributor to the group retention trend.

The concepts of "original group" and "percent retained" as used by LOGPRO (see LOGPRO Technical Analysis, page 1) are carried over to LPFILE in a modified fashion. The "original group" is a subset of those hired in the past few (e.g., 5) years. The "percent retained" figures are obtained from a cumulative distribution which compares the number of employees in the group who could have served for at least X years with those who actually have served at least X years. Such a distribution could conceivably be carried down to number of days of service (e.g., every member of the group would have at least one day of service while only those hired five years ago have five years of service).

Calculation of this cumulative distribution is possible with the knowledge of two dates:

(1) The date of hire for each member of the employee group under study; and

(2) The date of separation for every member who left the group.

---

[1] Users of LPFILE should also refer to documentation for LOGPRO

To perform its calculations, LPFILE requires a data file consisting of four data elements:

(1) Occupation or series code;

(2) Grade at hire;

(3) Date of hire (month and year); and

(4) Date of separation (month and year), if any.

When these data items are being collected, it will be necessary to record another piece of information: an employee identification number. This item is needed to facilitate the bringing together of the other items of data.

The conversion of such a data file into a computer file is discussed in the LPFILE Operation Manual (pages 1-2).

LPFILE also requires the input of a "file ending date" This is the closing date of the data file and of the time period under study.

Statistical Analysis

From the inputted data file and the file ending date, LPFILE sets up two one-dimensional arrays each of which contains one value for each employee in the file. The first array (T1) consists of values of the actual time on board for each employee. The second (T2) consists of the elapsed time between the employee date of hire and the file ending date. For those employees who have not separated, these two values would be equal.

These values are then used to determine the X, Y pairs needed for the log-probability regression analysis.

Quantitatively, this process is:

Let:

Ni = Employee i

DOHi = Date of hire for employee i
DOSi = Date of separation for employee i [2]
DOF = File ending date

---

[2] If an employee has not separated, both the month and year of the date of separation have the value zero.

Then:

T2 (Ni) = DOF - DOHi

If employee Ni has separated from the original group, then:

T1 (Ni) = DOSi - DOHi

If not, then:

T1 (Ni) = T2 (Ni)

(These "length of time" calculations are performed by the subroutine DATE which does a straightforward subtraction of two inputted dates.)

Each value in T1 and T2 is then converted into its base 10 logarithm.

The next steps form a cycle which calculates the X, Y pairs. A cycle begins with the finding of a maximum value from the T2 vector (TMAX). Each cycle produces a new TMAX and this value gets smaller each time through the cycle (as old values of TMAX are marked as such).

Let:

NP = the number of employees who could have possibly survived to at least time TMAX.

NR = the number of employees who did survive to at least time TMAX (i.e., were retained).

Both these values serve as counters and both are initialized at zero at the start of a cycle.

For every value in the T2 vector which is greater than or equal to TMAX, one is added to the NP counter. If, in these cases, the associated value in T1 is also greater than or equal to TMAX, one is added to the NR counter.

At the end of each cycle, a "percent retained" value, PR, is calculated using the summed values NP and NR:

$$PR = \frac{NR}{NP}$$

Then an X, Y pair is recorded for the log-probability regression analysis with (using ANDPX - See LOGPRO Technical Analysis, page 7):

X = TMAX

Y = ANDPX (PR) [3]

---

[3] ANDPX acts as a function whose value is given by the computation of the approximation contained in the subprogram.

A new cycle is then begun. This process continues until all possible values of TMAX have been used. When this point is reached, all the recorded X, Y pairs are used to calculate a log-probability equation and a standard deviation. These results are output to the user.

Data Groupings

The operator is given the option to run LPFILE using one or both of the first two data items in the file: occupation code and grade at hire. It can also be run for the entire input file.

This option allows the user to construct data files which contain records on employees in different occupations. (All employees in one occupation should be placed in the same data file.) However, in these cases, separate runs should be made for each occupation.

The ability to calculate log-probability equations by grade at hire is useful for determining whether differences in entrance levels lead to differences in retention patterns.

The "entire file" option can be used when there is one occupation per data file.

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LPFILE

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

"LPFILE" IS A FORTRAN IV PROGRAM WHICH CALCULATES A LOG-PROBABILITY EQUATION FOR UNGROUPED, NON-LONGITUDINAL RETENTION DATA. IT REQUIRES THREE SUBPROGRAMS: "ANDXP", "ANDPX" AND "DATE".

DATA REQUIREMENTS:

DATA ARE ENTERED INTO THE "LPFILE" PROGRAM FROM PREVIOUSLY-STORED DATA FILES. THESE FILES CONTAIN ONE RECORD PER EMPLOYEE AND CONSIST OF FOUR SPECIFIC DATA ELEMENTS PER RECORD:

(1) OCCUPATION CODE;

(2) GRADE AT HIRE;

(3) DATE OF HIRE (MONTH AND YEAR); AND

(4) DATE OF SEPARATION (MONTH AND YEAR), IF ANY.

THE COLLECTION PHASE FOR THE DATA FILES REQUIRES AN ADDITIONAL ITEM OF DATA: AN EMPLOYEE IDENTIFICATION NUMBER (E.G., SOCIAL SECURITY ACCOUNT NUMBER). THIS ITEM IS NECESSARY SINCE THE OTHER ITEMS ARE MOST LIKELY FOUND ON SEPARATE PIECES OF PAPER.

A SINGLE DATA FILE MAY CONSIST OF (A) ALL EMPLOYEES OR (B) ALL EMPLOYEES IN A GIVEN OCCUPATION OR (C) ALL EMPLOYEES IN A FEW GIVEN OCCUPATIONS.

A SAMPLE COLLECTION FORM MIGHT BE:

EMPLOYEE NUMBER	OCCUPATION CODE	GRADE AT HIRE	DATE OF HIRE	DATE OF SEPARATION
1	201	5	10/70	4/74
2	212	7	3/72	NONE
3	212	5	1/71	2/72
4	201	7	5/74	NONE
5	201	7	3/73	10/73
ETC.				

A FORM SUCH AS THIS MAY BE CONVERTED TO A COMPUTER FILE IN TWO WAYS. THE CHOICE OF METHOD WILL DEPEND ON WHAT THE USER'S TIME-SHARING SYSTEM WILL ACCEPT. IN ANY CASE, IF AN EMPLOYEE HAS NOT SEPARATED, BOTH THE MONTH AND YEAR OF SEPARATION ARE GIVEN THE VALUE ZERO.

IN ADDITION, THE DATA ITEMS ARE ALWAYS PLACED IN THE FILE IN THE FOLLOWING ORDER: OCCUPATION CODE, GRADE AT HIRE, MONTH OF HIRE, YEAR OF HIRE, MONTH OF SEPARATION, YEAR OF SEPARATION.

FORMATTED DATA FILE -- IF THE DATA FILE IS IN THIS FORM, THEN ALL DATA ITEMS FOLLOW EACH OTHER WITH NO SEPARATION CHARACTER BETWEEN THEM. HOWEVER, ALL SPACES CALLED FOR IN THE FORTRAN

"FORMAT" STATEMENT MUST BE ACCOUNTED FOR IN EACH LINE OF THE FILE (ONE LINE OF FILE = ONE EMPLOYEE).

AT PRESENT, THE INPUT FORMAT STATEMENT FOR DATA FILES IN "LPFILE" IS:

```
FORMAT(I5,5I2)
```

THIS FORMAT RESERVES FIVE SPACES FOR OCCUPATION CODE AND TWO SPACES EACH FOR GRADE AT HIRE, MONTH AND YEAR OF HIRE, AND MONTH AND YEAR OF SEPARATION. ALL VALUES ARE IN INTEGER FORM WHICH MEANS THERE ARE NO DECIMAL POINTS.

IN THIS CASE, THE SAMPLE FORM ABOVE WOULD BE CONVERTED TO A COMPUTER FILE IN THE FOLLOWING WAY:

```
002010510700474
002120703720000
002120501710272
002010705740000
002010703731073
ETC.
```

NOTE - IN MOST SYSTEMS, LEADING ZEROES MAY BE REPLACED BY BLANKS WITHOUT AFFECTING THE OPERATION OF THE PROGRAM.

UNFORMATTED DATA FILE -- IF THE DATA FILE IS IN THIS FORM, THEN THE DATA ITEMS IN THE FILE ARE SEPARATED BY COMMAS. IN THIS CASE, THE SAMPLE COLLECTION FORM WOULD BE CONVERTED TO A COMPUTER FILE IN THE FOLLOWING WAY:

```
201,5,10,70,4,74
212,7,3,72,0,0
212,5,1,71,2,72
201,7,5,74,0,0
201,7,3,73,10,73
ETC.
```

DEPENDING UPON THE COMPUTER SYSTEM USED, SUCH A FILE MAY BE READ IN BY A NUMBERED FORMAT STATEMENT (SUCH AS THE ONE ABOVE) OR IT MAY BE READ IN USING AN "*" IN PLACE OF A FORMAT STATEMENT NUMBER.

EXECUTION COMMANDS:

TO BEGIN EXECUTION OF THE "LPFILE" PROGRAMMING SEQUENCE, AN OPERATOR WILL PERFORM A CHAIN OF EXECUTION COMMANDS. THE ACTUAL FORM OF THESE COMMANDS WILL DEPEND ON THE TIME-SHARING SYSTEM BEING USED. IN GENERAL, THESE COMMANDS WILL PERFORM THE FOLLOWING OPERATIONS:

- CALL UP "LPFILE" AND ITS THREE SUBPROGRAMS "ANDXP", "ANDPX" AND "DATE" AND TRANSLATE THEM INTO MACHINE

LANGUAGE. THIS IS THE COMPILATION PHASE.

- LOAD THE COMPILED PROGRAM AND SUBPROGRAMS INTO THE CENTRAL PROCESSING AREA AND START PROGRAM RUN. THIS IS THE EXECUTION PHASE.

DATA ENTRY:

(NOTE THAT ALL OPERATOR-ENTERED RESPONSES TO COMPUTER-WRITTEN COMMANDS ARE FOLLOWED BY A CARRIAGE RETURN.)

THE "LPFILE" PROGRAMMING SEQUENCE BEGINS WITH THE FOLLOWING COMMAND:

```
*ENTER FILE ENDING DATE (MONTH, YEAR)
```

AT THIS POINT, THE OPERATOR ENTERS THE MONTH AND YEAR OF THE FILE ENDING DATE. THIS IS THE CLOSING DATE OF THE DATA FILE AND OF THE TIME PERIOD UNDER STUDY. THE DATE IS ENTERED IN THE FORM MONTH, YEAR (E.G., 6,75).

NEXT, THE COMPUTER ASKS:

```
*ENTER THE NUMBER OF EMPLOYEES IN THE FILE
```

THE OPERATOR ENTERS THE NUMBER OF EMPLOYEES FOR WHICH THERE ARE RECORDS IN THE FILE.

NEXT:

```
*ENTER THE NAME OF YOUR TURNOVER DATA FILE
*(MUST BE LESS THAN OR EQUAL TO 5 CHARACTERS)
```

AS EACH DATA FILE IS ENTERED INTO COMPUTER STORAGE, IT IS GIVEN A NAME TO IDENTIFY IT. IT IS THIS NAME THAT IS ENTERED HERE.

AT PRESENT, THE FORMAT FOR THE INPUT OF THE FILE NAME IS "A5". IT WILL ACCEPT A NAME UP TO FIVE CHARACTERS IN LENGTH. HOWEVER, SOME COMPUTER SYSTEMS MIGHT REQUIRE A DIFFERENT FORMAT.

IN ANY CASE, THE FILE NAME MUST BEGIN WITH AN ALPHABETIC CHARACTER.

NEXT, THE COMPUTER ASKS:

```
*DO YOU WISH BREAKDOWN BY:
*       (1) OCCUPATION
*       (2) GRADE
*       (3) BOTH OR
*       (4) NONE.  RUN WHOLE FILE
*       (ANS 1, 2, 3 OR 4)
```

"LPFILE" CAN BE USED TO CALCULATE LOG-PROBABILITY EQUATIONS BASED ON OCCUPATION, GRADE AT HIRE OR BOTH. IT CAN ALSO BE

* = COMPUTER WRITTEN

USED TO CALCULATE ONE LOG-PROBABILITY EQUATION FOR THE ENTIRE INPUT FILE. THIS OPTION WILL BE USED MOST OFTEN WHEN FILES ARE SEPARATED BY OCCUPATION. HOWEVER, IF ONE FILE CONTAINS MORE THAN ONE OCCUPATION, SEPARATE RUNS SHOULD BE MADE.

IF OPTION "1" IS CHOSEN, THE COMPUTER ASKS:

*ENTER DESIRED OCCUPATION CODE

THE OPERATOR ENTERS THE NUMERICAL CODE OF THE OCCUPATION DESIRED (E.G., 201). THE COMPUTER WILL THEN SEARCH THE INPUT FILE AND PULL OUT ONLY THOSE RECORDS OF EMPLOYEES WITH THAT OCCUPATION.

IF OPTION "2" IS CHOSEN, THE COMPUTER ASKS:

*ENTER DESIRED GRADE

THE OPERATOR ENTERS THE NUMERICAL VALUE OF THE GRADE DESIRED (E.G., 7). UNDER THIS OPTION, ONLY THE RECORDS OF EMPLOYEES WITH THAT PARTICULAR GRADE AT HIRE ARE PULLED OUT.

IF OPTION "3" IS CHOSEN, THE COMPUTER TYPES OUT BOTH OF THE COMMANDS ABOVE AND THE USER ENTERS THE APPROPRIATE NUMERICAL VALUES. IN THIS CASE, THE COMPUTER SEARCH IS FOR RECORDS WITH THE DESIRED OCCUPATION AND GRADE COMBINATION.

UNDER OPTION "4", NO FURTHER QUESTIONS ARE ASKED AND THE COMPUTER USES THE WHOLE FILE FOR ANALYSIS PURPOSES.

ANALYSIS OUTPUTS:

THE COMPUTER THEN CALCULATES A LOG-PROBABILITY EQUATION BASED ON THE INPUT PARAMETERS. A STANDARD DEVIATION IS ALSO CALCULATED. THE RESULTS ARE THEN OUTPUT TO THE USER IN A LABELED FORM.

UNDER OPTION "1", THE LABEL IS IN THE FORM:

*FOR OCCUPATION: XXX

UNDER OPTION "2", IT IS:

*FOR GRADE: XX

OPTION "3" COMBINES THE FIRST TWO LABELS:

*FOR OCCUPATION: XXX
*FOR GRADE: XX

UNDER OPTION "4", THE LABEL IS:

*FOR THE ENTIRE FILE:

THEN THE RESULTS ARE PRINTED OUT IN THE FOLLOWING FORMAT:

```
*THE L-P EQUATION IS:
*Y =   0.12345 + (-1.23456)X
*
*AND STANDARD DEVIATION IS:   0.12345
```

NEXT,

```
*WRITE OUT ACTUAL AND CURVE VALUES? (Y OR N)
```

AT THIS POINT, THE OPERATOR HAS THE OPTION TO WRITE OUT THE X,Y PAIRS WHICH WERE USED TO CALCULATE THE LOG-PROBABILITY EQUATION.

IF "Y" (OR YES) IS INPUT, THE ACTUAL AND CURVE VALUES ARE PRINTED OUT UNDER THE FOLLOWING HEADINGS:

```
*THE RETENTION VALUES ARE:
*
*     L.O.S.   -PERCENT RETAINED-
*     (YEARS)  -ACTUAL-    -CURVE-
```

RECYCLING:

NEXT, THE COMPUTER ASKS:

```
*AGAIN WITH THE SAME FILE? (Y OR N)
```

IF THERE ARE MORE OPERATIONS TO BE RUN FOR THE SAME INPUT FILE (SUCH AS FINDING A LOG-PROBABILITY EQUATION FOR A DIFFERENT OCCUPATION), THEN THE ANSWER TO THIS QUESTION IS "Y". IF "Y" IS ENTERED, THE COMPUTER RECYCLES TO THE POINT WHERE A TYPE OF BREAKDOWN IS REQUESTED (SEE PAGE 3).

IF THERE ARE NO MORE OPERATIONS TO BE RUN ON THIS SAME FILE, THE ANSWER IS "N". IF "N" IS ENTERED, THE COMPUTER TYPES:

```
*AGAIN WITH ANOTHER FILE? (Y OR N)
```

IF THERE IS ANOTHER FILE FOR WHICH ANALYSIS IS DESIRED, THE ANSWER TO THIS QUESTION IS "Y". THE COMPUTER THEN RECYCLES TO THE BEGINNING OF THE PROGRAM.

IF THERE ARE NO MORE FILES TO BE ANALYZED, THEN THE ANSWER IS "N". THE COMPUTER THEN TERMINATES THE RUN OF "LPFILE".

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LPFILE

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

LPFILE

THIS PROGRAM PERFORMS LOG-PROBABILITY ANALYSIS ON UNGROUPED, NON-LONGITUDINAL RETENTION DATA.

DATA ARE ENTERED INTO THE PROGRAM FROM PREVIOUSLY-STORED EMPLOYEE FILES CONSISTING OF FOUR DATA ELEMENTS PER EMPLOYEE (OCCUPATION CODE, GRADE AT HIRE, DATE OF HIRE, DATE OF SEPARATION, IF ANY). ONE FILE RECORD EQUALS ONE EMPLOYEE.

LOG-PROBABILITY EQUATIONS MAY BE CALCULATED BY OCCUPATION AND/OR GRADE AT HIRE.

REQUIRED SUBPROGRAMS: ANDPX, ANDXP, DATE

```
      DIMENSION IT(6),T1(500),T2(500),XX(500),YY(500),YA(500)
20    WRITE(5,30)
30    FORMAT(//////,1X,´ENTER FILE ENDING DATE (MONTH, YEAR)´)
      READ(5,*)MOF,MYF
50    FORMAT(2I2)
60    WRITE(5,70)
70    FORMAT(/,1X,´ENTER THE NUMBER OF EMPLOYEES IN THE FILE´)
      READ(5,90)N
90    FORMAT(I4)
      WRITE(5,110)
110   FORMAT(/,1X,´ENTER THE NAME OF YOUR TURNOVER DATA FILE´
      1  ,/,1X,´(MUST BE LESS THAN OR EQUAL TO 5 CHARACTERS)´)
      READ(5,140)NAME
140   FORMAT(A5)
150   WRITE(5,160)
160   FORMAT(/,1X,´DO YOU WISH BREAKDOWN BY:´,/,8X,´(1) OCC´
      1  ´UPATION´,/,8X,´(2) GRADE´,/,8X,´(3) BOTH OR´,/,
      2  8X,´(4) NONE. RUN WHOLE FILE´,/,8X,´(ANS 1, 2, 3´
      3  ´ OR 4)´)
      READ(5,210)IBR
210   FORMAT(I2)
      IF (IBR.EQ.4)GO TO 330
      IF(IBR.EQ.2)GO TO 290
      WRITE(5,250)
250   FORMAT(/,1X,´ENTER DESIRED OCCUPATION CODE´)
      READ(5,270)NOC
270   FORMAT(I5)
      IF (IBR.EQ.1)GO TO 330
290   WRITE(5,300)
300   FORMAT(/,1X,´ENTER DESIRED GRADE´)
      READ(5,320)NGR
320   FORMAT(I2)
330   CALL IFILE(20,NAME)
      KI=0
      DO 510 I=1,N
350   READ(20,355)(IT(J),J=1,6)
355   FORMAT(I5,5I2)
      IF(IBR.EQ.4) GO TO 410
      IF(IBR.EQ.2)GO TO 400
      IF(IT(1).NE.NOC)GO TO 510
      IF(IBR.EQ.1) GO TO 410
400   IF(IT(2).NE.NGR) GO TO 510
410   MOH=IT(3)
      MYH=IT(4)
      MOS=IT(5)
      MYS=IT(6)
      IF(MOS.EQ.0)GO TO 480
      CALL DATE(MOS,MYS,MOH,MYH,DA)
      IF(DA.EQ.0)GO TO 510
480   CALL DATE(MOF,MYF,MOH,MYH,DT)
      IF(DT.EQ.0)GO TO 510
      KI=KI+1
      T2(KI)=DT
      IF(MOS.EQ.0)T1(KI)=DT
      IF(MOS.NE.0)T1(KI)=DA
```

```
510     CONTINUE
        TMAX=ALOG10(T2(1))
        TMIN=TMAX
        DO 580 I=1,KI
        T1(I)=ALOG10(T1(I))
        T2(I)=ALOG10(T2(I))
        IF(T2(I).LT.TMIN)TMIN=T2(I)
580     IF(T2(I).GT.TMAX)TMAX=T2(I)
        SUMN=0.0
        SUMX=0.0
        SUMY=0.0
        SUMXY=0.0
        SUMX2=0.0
        J=0
        M=0
660     ENPOS=0.0
        ENRET=0.0
        DO 740 K=1,KI
        IF(T2(K).LT.TMAX)GO TO 740
        ENPOS=ENPOS+1.
        IF(T1(K).GE.TMAX)ENRET=ENRET+1.
        IF(T2(K).NE.9.)J=J+1
        T2(K)=9.
740     CONTINUE
        RET=ENRET/ENPOS
        IF((RET.EQ.1.).OR.(RET.EQ.0.0))GO TO 890
        Y=ANDPX(RET)
        X=TMAX
        M=M+1
        XX(M)=10.**X
        YY(M)=RET
        NO=ENPOS
C       DO 880 MM=1,NO
        SUMN=SUMN+1.
        SUMX=SUMX+X
        SUMY=SUMY+Y
        SUMXY=SUMXY+X*Y
880     SUMX2=SUMX2+X**2
890     IF(J.EQ.KI)GO TO 940
        TMAX=TMIN
        DO 920 L=1,KI
920     IF((T2(L).NE.9.).AND.(T2(L).GT.TMAX))TMAX=T2(L)
        IF(J.LT.KI) GO TO 660
940     EN=SUMN
        S1=EN*SUMX2-SUMX*SUMX
        S2=EN*SUMXY-SUMX*SUMY
        B=S2/S1
        A=SUMY/EN-(B*SUMX)/EN
        SUMD2=0.0
        DO 1020 LL=1,M
        YA(LL)=ANDXP(A+B*ALOG10(XX(LL)))
1020    SUMD2=SUMD2+(YY(LL)-YA(LL))**2
        EM=M
        SD=SQRT(SUMD2/(EM-1.))
        WRITE(5,1060)
```

```
1060     FORMAT(////////)
         IF(IBR.EQ.1.OR.IBR.EQ.3)WRITE(5,1080)NOC
1080     FORMAT(1X,'FOR OCCUPATION: ',I5)
         IF((IBR.EQ.2).OR.(IBR.EQ.3))WRITE(5,1100)NGR
1100     FORMAT(1X,'FOR GRADE: ',I2)
         IF(IBR.EQ.4)WRITE(5,1120)
1120     FORMAT(1X,'FOR THE ENTIRE FILE:')
         WRITE(5,1240)A,B
1240     FORMAT(/,1X,'THE L-P EQUATION IS: ',/,1X,'Y = ',F9.5,' + ('
         1  ,F9.5,')X')
         WRITE(5,1270)SD
1270     FORMAT(/,1X,'AND THE STANDARD DEVIATION IS: ',F9.5)
         WRITE(5,1127)
1127     FORMAT(/,1X,'WRITE OUT ACTUAL AND CURVE VALUES? (Y OR N)')
         READ(5,1274)KWR
         IF(KWR.EQ.'N')GO TO 1271
         WRITE(5,1140)
1140     FORMAT(/,1X'THE RETENTION VALUES ARE:')
         WRITE(5,1160)
1160     FORMAT(/,6X,'L.O.S.',3X,'-PERCENT RETAINED-')
         WRITE(5,1180)
1180     FORMAT(6X,'(YEARS)',2X,'-ACTUAL-',3X,'-CURVE-',/)
         MN=M
         DO 1220 MA=1,M
         WRITE(5,1210)XX(MN),YY(MN),YA(MN)
1210     FORMAT(6X,F7.2,2X,F8.3,3X,F7.3)
         MN=MN-1
1220     CONTINUE
1271     WRITE(5,1272)
1272     FORMAT(///,1X,'AGAIN WITH SAME FILE? (Y OR N)')
         READ(5,1274)NAG
1274     FORMAT(A2)
         IF(NAG.EQ.'Y')REWIND 20
         IF(NAG.EQ.'Y')GO TO 150
         WRITE(5,1277)
1277     FORMAT(/,1X,'AGAIN WITH ANOTHER FILE? (Y OR N)')
         READ(5,1274)KAG
         IF(KAG.EQ.'Y')GO TO 20
         WRITE(5,1060)
         STOP
         END
```

```
      FUNCTION ANDPX(P)
C     ACCUMULATIVE NORMAL DISTRIBUTION FUNCTION.
C     APPROXIMATES X (+ OR - .000004) FROM P.
      DATA A0,A1,A2,B1,B2,B3/2.515517,.802853,.010328,1.432788,
     1  .189269,.001308/
      IF(P.GE.1.) X1=4.
      IF(P.GE.1.) GO TO 380
      IF(P.GT.0.0005) GO TO 70
      X1=-4.
      GO TO 380
70    IF (P.LT.0.5) GO TO 110
      E = SQRT(ALOG (1./(1.-P)**2))
      X1 = E-((A2*E+A1)*E+A0)/(((B3*E+B2)*E+B1)*E+1.)
      GO TO 140
110   P = 1.-P
      E = SQRT(ALOG(1./(1.-P)**2))
      X1 = -1.*(E-((A2*E+A1)*E+A0)/(((B3*E+B2)*E+B1)*E+1.))
      P = 1.-P
140   AX1 = X1 +0.0005
      AX2 = X1 -0.0005
      AP1 = ANDXP(AX1)
      AP2 = ANDXP(AX2)
      AI = (P-AP2)/(AP1-AP2)
      BX1 = AX2+AI*(AX1-AX2)
      D1 = P-ANDXP(BX1)
      IF (D1.LT.0.000001) GO TO 290
      AX3 = BX1+1.1*D1
      AX4 = BX1-1.1*D1
      AP3 = ANDXP(AX3)
      AP4 = ANDXP(AX4)
      AII = (P-AP4)/(AP3-AP4)
      BX1 = AX4+AII*(AX3-AX4)
      D1 = P-ANDXP(BX1)
290   IF (D1.LE.0.0000005) GO TO 370
      IF (D1.GT.0.0000005) GO TO 340
      BX1 = BX1+0.0000002
      D1 = ANDXP(BX1)
      GO TO 290
340   BX1 = BX1-0.0000002
      D1 = ANDXP(BX1)
      GO TO 290
370   ANDPX = BX1
      GO TO 390
380   ANDPX=X1
390   RETURN
      END
      FUNCTION ANDXP(X)
C     ACCUMULATIVE NORMAL DISTRIBUTION FUNCTION. APPROXIMATES P
C     FROM X (TO + OR - .0000001). (NBS-55, P. 932)
      X1 = ABS(X)
      T = 1./(1.+.2316419*X1)
      ANDXP = 1.-.3989423*EXP(-(X1**2)/2.)*(.3193815*T-.3565638*T**2
     1  +1.781478*T**3-1.821256*T**4+1.330274*T**5)
      IF (X.LT.0.0) ANDXP = 1.-ANDXP
      RETURN
      END
```

```
SUBROUTINE DATE(MOI,MYI,MOF,MYF,DAT)
NY=MYI-MYF
NM=MOI-MOF
EN=NM
EN=EN/12.
ENY=NY
DAT=ENY+EN
RETURN
END
```

APPENDIX E

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LPTEST

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LPTEST

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

LPTEST TECHNICAL ANALYSIS[1]

Introduction

The computer terminal program LPTEST was developed to enable users to:

(1) Compare sets of retention data to determine whether or not they may be grouped together; and

(2) Use the results of these comparisons as input into the log-probability analysis program (LOGPRO).

LPTEST and its three subprograms (ANDXP, ANDPX, and FTEST) together form a comprehensive, self-contained unit which completes all the necessary statistics required to compare retention groups. This means that a user of this set of programs need not have an extensive statistical background in order to successfully utilize and evaluate its results.

The form of the input data, the statistical transformations, and the basic methodological assumptions are the same as those for LOGPRO. (See LOGPRO Technical Analysis, pages 1-8.)

Data Groupings

The basic grouping of data for LPTEST will be by individual occupation and the basic comparison test will be to determine what occupations can be grouped together for log-probability analysis purposes. However, it is also possible to use groupings which are the sum of two or more occupations. In this case, it will generally have previously been determined that these occupations are "compatible." Summed groupings of occupations may be tested against either other summed groupings or individual occupations.

In addition, groupings can be based on other factors which might be of interest to the user. Some of these factors are sex, minority status, veterans preference, grade at hire, etc. Since retention rates may vary greatly among occupations, comparisons based on these factors should be made within the same occupations.

1/ Since the results of this program will be used to determine the input to LOGPRO, users of LPTEST should also have the documentation for LOGPRO. Thus, this Technical Analysis refers to the Technical Analysis for LOGPRO rather than re-describing common aspects of the two processes.

Statistical Analysis

LPTEST allows for the initial input of retention data for up to twenty occupations or groups. The operator may select any or all of these groups for retention trend comparisons. These comparisons are made using an analysis of variance technique.

Calculation of Variance - After selecting out the (NG) groups specified by the user (each group having NP retention points), LPTEST calculates two variances which are used to determine a value of the F-statistic. These are:

(1) The variance among the tested groups; and

(2) The variance within the tested groups.

The F-statistic is then calculated using the formula:

$$F = \frac{\text{Variance among groups}}{\text{Variance within groups}}$$

Where:

V_1 = degrees of freedom associated with the variance among the groups

= NP-1

V_2 = degrees of freedom associated with the variance within the groups

= (NP x NG) - 1

Test of Significance - The probability of chance occurrence of the calculated F-statistic is directly determined by the subroutine FTEST. This subprogram uses the following approximation[2] for X (the number of standard deviations from the mean of a normal curve):

$$X = \frac{F^{1/3}\left(1 - \frac{2}{9V_2}\right) - \left(1 - \frac{2}{9V_1}\right)}{\sqrt{\frac{2}{9V_1} + F^{2/3}\left(\frac{2}{9V_2}\right)}}$$

---

[2] Abramowitz, M. and Stegun, I.A., Eds., Handbook of Mathematical Functions, AMS55, 9th, National Bureau of Standards, 26.6.15, p. 947.

The subprogram then uses ANDXP to calculate the accumulative normal distribution function value (Q) associated with X. Then, the probability of chance occurrence (P) is:

$$P = 1 - Q$$

P-values of 0.05 or less are considered to be statistically significant. The value of P determines whether or not the tested groups may be combined.

Log-Probability Analysis - A log-probability equation is calculated for each of the tested groups. In addition, if the groups are found to be compatible, their retention data are summed and a log-probability equation is calculated for the total group. The statistical techniques used in these calculations are explained in the LOGPRO Technical Analysis (pages 6-8).

Analysis Outputs

The form of the final output of LPTEST is labeled with the codes of the occupations or groups tested. The output consists of:

(1) A statement of grouping which tells whether or not the tested groups may be combined;

(2) A table of data showing, for each group tested:

  (a) Occupation code,
  (b) Number in each starting group, and
  (c) A- and B-values of each subgroup's log-probability equation; and

(3) When the tested groups are compatible, the A- and B-values of the combined group's log-probability equation.

There are options provided which allow an operator to (a) make further tests using groups whose data have already been entered or (b) enter and test a different set of groups.

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LPTEST

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

"LPTEST" IS A FORTRAN IV PROGRAM WHICH COMPARES THE RETENTION TRENDS OF TWO OR MORE OCCUPATIONS TO DETERMINE IF THEY CAN BE GROUPED TOGETHER FOR INPUT INTO "LOGPRO" (THE LOG-PROBABILITY ANALYSIS PROGRAM). "LPTEST" REQUIRES THREE SUBPROGRAMS: "ANDPX", "ANDXP", AND "FTEST".

DATA REQUIREMENTS:

THE "LPTEST" PROGRAMMING SEQUENCE IS DESIGNED TO ANALYZE AND COMPARE TWO OR MORE SETS OF LONGITUDINAL "RETENTION-OVER-TIME" DATA. THE RULES OF DATA COLLECTION FOR "LPTEST" ARE THE SAME AS THOSE FOR "LOGPRO". (SEE "LOGPRO" TECHNICAL ANALYSIS OR INSTRUCTION MANUAL.)

GENERALLY, THE DATA SETS INVOLVED WILL REPRESENT DIFFERENT OCCUPATIONS. HOWEVER, IT IS ALSO POSSIBLE TO COMPARE GROUPS WHICH ARE EITHER SUBGROUPS OF ONE OCCUPATION OR ALREADY-GROUPED OCCUPATIONS. (SEE GROUPING OCCUPATIONS, PAGE 8).

HYPOTHETICAL DATA SETS:

THE FOLLOWING ARE HYPOTHETICAL DATA SETS OF THE TYPE REQUIRED FOR THE "LPTEST" PROGRAMMING SEQUENCE:

> SUPPOSE THAT DURING FY 1972 AN ORGANIZATION HIRED 110 CLERK-TYPISTS (GS-322), 125 SECRETARIES (GS-318), AND 270 PERSONNEL MANAGEMENT SPECIALISTS (GS-201). SUPPOSE THAT THESE HIRES WERE MADE EVENLY THROUGHOUT THE FISCAL YEAR. AND SUPPOSE THAT THE FOLLOWING RETENTION VALUES WERE RECORDED OVER THE NEXT THREE FISCAL YEARS:

FOR 322:

AT THE END OF FY:	THE NUMBER RETAINED WAS:
1973	58 (OR 52.73%)
1974	44 (OR 40.00%)
1975	36 (OR 32.73%)

FOR 318:

AT THE END OF FY:	THE NUMBER RETAINED WAS:
1973	77 (OR 61.60%)
1974	62 (OR 49.60%)
1975	53 (OR 42.40%)

FOR 201:

AT THE END OF FY:	THE NUMBER RETAINED WAS:
1973	225 (OR 83.33%)
1974	201 (OR 74.44%)
1975	190 (OR 70.37%)

THE AVERAGING FACTOR FOR THESE DATA SETS IS 0.5 YEARS. THESE SETS OF DATA WILL BE REFERRED TO THROUGHOUT THIS MANUAL.

* = COMPUTER WRITTEN

EXECUTION COMMANDS:

TO BEGIN EXECUTION OF THE "LPTEST" PROGRAMMING SEQUENCE, AN OPERATOR WILL PERFORM A CHAIN OF EXECUTION COMMANDS. THE ACTUAL FORM OF THESE COMMANDS WILL DEPEND ON THE TIME-SHARING SYSTEM BEING USED. IN GENERAL, THESE COMMANDS WILL PERFORM THE FOLLOWING OPERATIONS:

- CALL UP "LPTEST" AND ITS THREE SUBPROGRAMS "ANDXP", "ANDPX" AND "FTEST" AND TRANSLATE THEM INTO MACHINE LANGUAGE. THIS IS THE COMPILATION PHASE.
- LOAD THE COMPILED PROGRAM AND SUBPROGRAMS INTO THE CENTRAL PROCESSING AREA AND START PROGRAM RUN. THIS IS THE EXECUTION PHASE.

DATA ENTRY:

(NOTE THAT ALL OPERATOR-ENTERED RESPONSES TO COMPUTER-WRITTEN COMMANDS ARE FOLLOWED BY A CARRIAGE RETURN.)

THE RUN OF THE "LPTEST" PROGRAMMING SEQUENCE BEGINS WITH THE PRINTOUT OF THE FOLLOWING INFORMATION:

```
*THIS PROGRAM ANALYZES AND COMPARES THE RETENTION TRENDS OF 2 OR MORE
*OCCUPATIONS TO DETERMINE WHETHER THEY CAN BE GROUPED TOGETHER FOR LOG-
*PROBABILITY ANALYSIS.
*
*FOR PURPOSES OF THIS PROGRAM:
*     (1) THE "X - VALUES" = LENGTH OF SERVICE COMPLETED; AND
*     (2) THE "Y - VALUES" = NUMBER (OR PERCENT) RETAINED AT TIME X.
*
*SEE INSTRUCTION MANUAL FOR FURTHER EXPLANATION OF DATA REQUIRED.
```

NEXT, THE COMPUTER ASKS:

```
*ENTER THE NUMBER OF RETENTION GROUPS TO BE COMPARED
```

THE DESIGN OF LPTEST ALLOWS THE USER TO ENTER UP TO TWENTY OCCUPATIONS AT THE BEGINNING OF A RUN AND THEN SELECT OUT WHICH SPECIFIC OCCUPATIONS ARE TO BE COMPARED DURING EACH RUN OF LPTEST'S ANALYSIS SEQUENCE. ANY SUBSET OF THE INPUTTED OCCUPATIONS MAY BE COMPARED REGARDLESS OF THE NUMBER IN THE SUBSET OR THE ORDER IN WHICH THE ORIGINAL OCCUPATIONS HAVE BEEN ENTERED.

FOR EXAMPLE, USING THE HYPOTHETICAL DATA, ALL THREE OF THE SAMPLE OCCUPATIONS CAN BE ENTERED INTO LPTEST AT ONCE. THEN THE RESPONSE TO THIS COMMAND WOULD BE "3". OR, IF THE OPERATOR SO DESIRES, ONLY SOME OF THE OCCUPATIONS FOR WHICH DATA ARE AVAILABLE NEED BE ENTERED. THUS, IF ONLY 322 AND 318 ARE TO BE ENTERED AND COMPARED, THEN THE RESPONSE TO THE COMMAND IS "2".

EACH OF THE ENTERED GROUPS SHOULD HAVE THE SAME NUMBER OF KNOWN RETENTION POINTS AT THE SAME TIME VALUES.

* = COMPUTER WRITTEN
! = OPERATOR ENTERED

THE COMPUTER THEN ASKS FOR EACH OCCUPATION CODE IN TURN AND THE OPERATOR ENTERS THE CORRECT CODES. FOR EXAMPLE, IF THE DATA FOR 322, 318 AND 201 ARE TO BE ENTERED, THE SEQUENCE IS AS FOLLOWS:

```
*ENTER THE OCCUPATION CODE FOR GROUP
*
*NO.  1:
!322
*
*NO.  2:
!318
*
*NO.  3:
!201
```

(IF THE GROUPS BEING COMPARED ARE EITHER SUBSETS OF ONE OCCUPATION OR ALREADY-GROUPED OCCUPATIONS, THEN EACH SUCH GROUP MUST BE GIVEN ITS OWN NUMERIC CODE OF UP TO FIVE DIGITS.)

NEXT, THE COMPUTER ASKS:

```
*ENTER THE NO. OF X-VALUES
```

THE INPUTTED VALUE AT THIS POINT IS THE NUMBER OF RETENTION POINTS FOR WHICH DATA ARE AVAILABLE. FOR EXAMPLE, IF THREE YEARS OF DATA ARE AVAILABLE, THE RESPONSE IS "3". (NOTE - THERE MUST BE AT LEAST TWO X-VALUES.)

NEXT, THE COMPUTER ASKS FOR EACH X-VALUE IN TURN. THE OPERATOR INPUTS EACH TIME VALUE (INCLUDING AN AVERAGING FACTOR, IF ANY). USING THE HYPOTHETICAL DATA (WHOSE AVERAGING FACTOR IS 0.5), THE PROPER X-VALUES FOR 1973, 1974 AND 1975 ARE 1.5, 2.5 AND 3.5, RESPECTIVELY. THE INPUT SEQUENCE IS AS FOLLOWS:

```
*ENTER X-VALUE
*
*NO.  1:
!1.5
*
*NO.  2:
!2.5
*
*NO.  3:
!3.5
```

FOLLOWING THESE ENTRIES, THE COMPUTER ASKS FOR THE NUMBER OF EMPLOYEES WHO COMPOSED THE ORIGINAL GROUPS OF HIRES FOR THE OCCUPATIONS BEING COMPARED. THESE ENTRIES ARE REQUESTED IN THE SAME ORDER AS THE OCCUPATION CODES WERE ENTERED ABOVE.

```
* = COMPUTER WRITTEN
! = OPERATOR ENTERED
```

FOR EXAMPLE, IF DATA FOR ALL OF THE OCCUPATIONS COMPRISING THE HYPOTHETICAL DATA SETS ARE TO BE ENTERED AT THE OUTSET, THEN THE STARTING GROUP DATA WOULD BE INPUT AS FOLLOWS:

```
*ENTER STARTING POPULATION (N) FOR GROUP
*
*NO.  1:  ( 322)
!110
*
*NO.  2:  ( 318)
!125
*
*NO.  3:  ( 201)
!270
```

NEXT,

```
*ARE Y-VALUES IN
*   (1) NUMBER OR
*   (2) PERCENT FORM?
*   (ANS 1 OR 2)
```

IF THE RETENTION DATA ARE IN THE FORM "NUMBER OF EMPLOYEES RETAINED FROM THE ORIGINAL GROUP", THEN THE RESPONSE TO THE ABOVE COMMAND IS "1".

ON THE OTHER HAND, IF THE DATA ARE IN THE FORM "PERCENT OF EMPLOYEES RETAINED FROM THE ORIGINAL GROUP", THEN THE RESPONSE IS "2".

IF THE Y-VALUES ARE IN "NUMBER" FORM, THEN THE COMPUTER TYPES:

```
*IN INTEGER FORM (NO DEC. PTS.)
*AND SEPARATED BY COMMAS, ENTER THE
*Y-VALUES CORRESPONDING TO X =
```

THE COMPUTER THEN INDIVIDUALLY PRINTS OUT EACH OF THE INPUTTED X-VALUES. THE OPERATOR THEN TYPES IN EACH Y-VALUE ASSOCIATED WITH THAT X-VALUE. THE Y-VALUES ARE ENTERED IN THE SAME ORDER AS THE OCCUPATION CODES WERE ENTERED ABOVE. THESE NUMBERS ARE ENTERED ON ONE LINE AND SEPARATED BY COMMAS. FOR EXAMPLE, USING THE HYPOTHETICAL DATA, IF THERE ARE THREE X-VALUES (1.5, 2.5 AND 3.5) AND THREE OCCUPATIONS (322, 318 AND 201, IN THAT ORDER), THEN THE INPUT SEQUENCE IS AS FOLLOWS:

```
* 1.5000:
!58,77,225
*
* 2.5000:
!44,62,201
*
```

* = COMPUTER WRITTEN
! = OPERATOR ENTERED

```
* 3.5000:
!36,53,190
```

IF THE Y-VALUES ARE IN "PERCENT" FORM, THEN THE COMPUTER TYPES:

```
*IN DECIMAL FORM, SEPARATED BY COMMAS,
*ENTER THE Y-VALUES CORRESPONDING TO X =
```

USING THE SAME PROCESS AS ABOVE, THE COMPUTER PRINTS OUT THE INDIVIDUAL X-VALUES. HOWEVER, THIS TIME THE OPERATOR ENTERS THE PERCENTAGE VALUES IN DECIMAL FORM. (E.G., 52.73% IS ENTERED AS .5273). THESE VALUES ARE ENTERED ON A SINGLE LINE, SEPARATED BY COMMAS, AND IN THE SAME ORDER AS THE OCCUPATION CODES WERE ENTERED ABOVE. FOR EXAMPLE, IF ALL THREE HYPOTHETICAL DATA SETS WERE BEING ENTERED (IN THE ORDER: 322, 318, 201), THE INPUT SEQUENCE WOULD BE:

```
* 1.5000:
!.5273,.6160,.8333
*
* 2.5000:
!.4000,.4960,.7444
*
* 3.5000:
!.3273,.4240,.7037
```

DURING THE NEXT SECTION OF DATA ENTRY, THE OPERATOR CHOOSES WHICH OF THE INPUTTED GROUPS ARE TO BE COMPARED. THE FIRST QUESTION ASKED BY THE COMPUTER IN THIS SECTION IS:

```
*DO YOU WISH TO TEST (1) ALL OR (2) SOME OF THESE GROUPS?
*(ANS 1 OR 2)
```

IF, AT THIS TIME, ALL OF THE INPUTTED GROUPS ARE TO BE COMPARED, THEN THE ANSWER TO THIS QUESTION IS "1". LPTEST THEN COMPARES THE RETENTION TRENDS OF ALL OF THE OCCUPATIONS WHICH HAVE BEEN ENTERED TO DETERMINE WHETHER OR NOT THE ENTIRE SET OF OCCUPATIONS MAY BE GROUPED TOGETHER.

IF THE OPERATOR WISHES TO SELECT OUT CERTAIN OF THE INPUTTED OCCUPATIONS FOR TESTING, THEN THE RESPONSE TO THIS QUESTION IS "2". THEN THE COMPUTER WILL ASK:

```
*HOW MANY GROUPS DO YOU WISH TO TEST?
```

THE OPERATOR THEN ENTERS THE NUMBER OF OCCUPATIONS THAT ARE TO BE TESTED AT THIS TIME. FOR EXAMPLE, USING THE HYPOTHETICAL DATA, IF THE OPERATOR WISHED TO COMPARE 322 AND 318, THEN THE RESPONSE TO THIS QUESTION IS "2".

* = COMPUTER WRITTEN

TO FIND WHICH OCCUPATIONS ARE TO BE TESTED, THE COMPUTER ASKS:

```
*WHICH ONES?  ENTER USING THE GROUP NUMBERS
*ESTABLISHED DURING THE ENTRY OF OCCUPATION CODES.
*(SEPARATE THESE VALUES WITH COMMAS.)
```

THE OPERATOR THEN ENTERS THE OCCUPATIONS TO BE TESTED USING THE NUMBERS ASSIGNED TO THE INPUTTED GROUPS BY LPTEST DURING THE ENTRY OF OCCUPATION CODES. FOR THE HYPOTHETICAL DATA, 322 IS GROUP #1, 318 IS GROUP #2 AND 201 IS GROUP #3. SO, FOR EXAMPLE, IF 322 AND 318 ARE TO BE COMPARED THEN THE OPERATOR WOULD ENTER "1,2". THESE VALUES ARE ENTERED ON ONE LINE AND SEPARATED BY COMMAS. THE NUMBER OF VALUES IN THE LINE MUST EQUAL THE NUMBER OF GROUPS TO BE TESTED.

ANALYSIS OUTPUTS:

AFTER ALL OF THE REQUIRED DATA HAVE BEEN ENTERED, THE COMPUTER THEN PERFORMS ALL OF THE ANALYSIS NECESSARY TO DETERMINE WHETHER OR NOT THE TESTED OCCUPATIONS CAN BE GROUPED TOGETHER. THE NEXT STEP IS THE PRINTOUT OF RESULTS.

THE "LPTEST ANALYSIS OUTPUT" IS A COMPLETELY LABELED PRINTOUT WHICH CONSISTS OF:

(1) A "STATEMENT OF GROUPING" WHICH TELLS THE OPERATOR WHETHER OR NOT THE TESTED OCCUPATIONS OR GROUPS MAY BE COMBINED;

(2) A TABLE OF "INDIVIDUAL SUBGROUP DATA" SHOWING, FOR EACH GROUP TESTED:
    (A) THE OCCUPATION CODE,
    (B) THE NUMBER IN THE STARTING GROUP, AND
    (C) THE A- AND B- VALUES OF EACH GROUP'S LOG-PROBABILITY EQUATION; AND

(3) IF THE TESTED OCCUPATIONS MAY BE COMBINED, THE A- AND B- VALUES OF THE LOG-PROBABILITY EQUATION FOR THE COMBINED (OR SUMMED) GROUP.

EACH ANALYSIS OUTPUT LISTS THE OCCUPATIONS OR GROUPS WHICH WERE TESTED DURING THE CURRENT RUN.

FOR EXAMPLE, IF ALL THREE OCCUPATIONS IN THE HYPOTHETICAL DATA SET WERE TESTED, THEN THE RESULT WOULD LOOK LIKE THIS:

* = COMPUTER WRITTEN

```
*          XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
*          X                                      X
*          X                                      X
*          X  -------LPTEST ANALYSIS OUTPUT-------  X
*          X                                      X
*          X                                      X
*          X  THE  3 OCCUPATIONS TESTED           X
*          X        322                           X
*          X        318                           X
*          X        201                           X
*          X  MAY NOT BE GROUPED TOGETHER.        X
*          X                                      X
*          X                                      X
*          X  INDIVIDUAL SUBGROUP DATA:           X
*          X                                      X
*          X                 --L - P EQUATION--   X
*          X      OCCN     N      (A)       (B)   X
*          X      ----     -      ---       ---   X
*          X                                      X
*          X       322    110  0.31322  -1.40587  X
*          X       318    125  0.52560  -1.32681  X
*          X       201    270  1.16436  -1.19286  X
*          X                                      X
*          X                                      X
*          XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

SINCE THESE THREE OCCUPATIONS CANNOT BE GROUPED, A LOGICAL NEXT STEP WOULD BE TO COMPARE 322 AND 318 (THE CLERICAL OCCUPATIONS). THE RESULT OF SUCH A COMPARISON WOULD BE:

```
*          XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
*          X                                      X
*          X                                      X
*          X  -------LPTEST ANALYSIS OUTPUT-------  X
*          X                                      X
*          X                                      X
*          X  THE  2 OCCUPATIONS TESTED           X
*          X        322                           X
*          X        318                           X
*          X  MAY BE GROUPED TOGETHER.            X
*          X                                      X
*          X                                      X
*          X  INDIVIDUAL SUBGROUP DATA:           X
*          X                                      X
*          X                 --L - P EQUATION--   X
*          X      OCCN     N      (A)       (B)   X
*          X      ----     -      ---       ---   X
*          X                                      X
*          X       322    110  0.31322  -1.40587  X
*          X       318    125  0.52560  -1.32681  X
```

* = COMPUTER WRITTEN

```
*             X                                       X
*             X                                       X
*             X  TOTAL GROUP EQUATION:                X
*             X                                       X
*             X      Y =  0.42320 -  1.35371X         X
*             X                                       X
*             X                                       X
*             XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

AS YOU CAN SEE, THIS OUTPUT CONTAINS ONLY DATA FOR 322 AND 318 AND, SINCE THESE TWO OCCUPATIONS WERE FOUND TO BE COMPATIBLE, THE "TOTAL GROUP EQUATION".

RECYCLING:

AFTER PRINTING THIS OUTPUT THE COMPUTER ASKS:

*DO YOU WISH TO TEST ANOTHER SUBSET OF THESE GROUPS? (Y OR N)

IF THE OPERATOR WISHES TO TEST ANOTHER SET OF GROUPS FROM THOSE ALREADY-INPUTTED, THEN THE RESPONSE TO THIS QUESTION IS "Y". WITH THIS RESPONSE THE COMPUTER RECYCLES TO THE POINT WHERE GROUP SELECTION TAKES PLACE.

IF THE RESPONSE IS "N", THEN THE COMPUTER ASKS:

*DO YOU WISH TO TEST ANOTHER SET OF GROUPS? (Y OR N)

IF THERE IS ANOTHER SET OF GROUPS OR OCCUPATIONS TO BE ENTERED AND TESTED, THEN THE RESPONSE TO THIS QUESTION IS "Y" AND THE PROGRAM WILL RECYCLE BACK TO THE BEGINNING. IF NOT, THE RESPONSE IS "N" AND THE PROGRAM RUN ENDS.

GROUPING OCCUPATIONS:

"COMPATIBLE" OCCUPATIONS MAY BE GROUPED TOGETHER AND INPUTTED INTO "LPTEST". THIS IS DONE BY SUMMING THE Y-VALUES (NUMBER RETAINED) FOR EACH X-VALUE. THE "STARTING GROUP" FIGURE WOULD BE THE SUM OF THE INDIVIDUAL STARTING GROUPS. IF PERCENTAGE VALUES ARE DESIRED, THE INDIVIDUAL X-VALUE SUMS CAN BE DIVIDED BY THE SUMMED STARTING GROUP FIGURE. FOR EXAMPLE, USING THE HYPOTHETICAL DATA, ASSUME THAT 322 AND 318 CAN BE GROUPED TOGETHER. THEN THE SUMS FOR THE INDIVIDUAL X-VALUES WOULD BE:

X-VALUE	Y-VALUES		SUM
1.5	58+77	=	135
2.5	44+62	=	106
3.5	36+53	=	89

THE STARTING GROUP SUM IS: 110+125 = 235.
THE PERCENTAGE VALUES ARE: 57.45%, 45.11%, 37.87%.

THESE GROUPED FIGURES CAN BE COMPARED WITH OTHER GROUPED FIGURES OR WITH OTHER INDIVIDUAL OCCUPATIONS.

THE RETENTION VALUES FOR GROUPS WHICH ARE SUBGROUPS OF ONE OCCUPATION ARE COLLECTED BY DIVIDING THE STARTING GROUP FOR THE OCCUPATION INTO THE DESIRED SUBGROUPS (E.G., MALE - FEMALE) AND FOLLOWING EACH SUBGROUP OVER TIME. THE COMBINED VALUES FOR THESE SUBGROUPS WOULD GIVE THE RETENTION PATTERN FOR THE WHOLE OCCUPATION.

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

LPTEST

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

246

LPTEST

THIS PROGRAM COMPARES TWO OR MORE SETS OF LONGITUDINAL RETENTION DATA TO DETERMINE WHETHER ANY OR ALL OF THEM MAY BE GROUPED TOGETHER FOR LOG-PROBABILITY ANALYSIS. IN MOST CASES, THE COMPARISONS WILL BE BETWEEN TWO OR MORE OCCUPATIONS ALTHOUGH OTHER FACTORS MAY BE USED.

LONGITUDINAL RETENTION DATA POINTS ARE ENTERED INTO THE PROGRAM BY THE USER.

REQUIRED SUBPROGRAMS: ANDPX, ANDXP, FTEST

```
      DIMENSION X(20),XLOG(20),IY(20,20),Y(20,20),YSD(20,20),NO(20)
     1  ,A(20),B(20),IOC(20),NXX(40),
     2  YPT(20,20),SMP(20),SMP2(20),NUM(20),YD(20,20)
      DATA (NXX(I),I=1,40)/40*'XX'/
      WRITE(5,5)
5     FORMAT(10(/),1X,'THIS PROGRAM ANALYZES AND COMPARES THE'
     1  ' RETENTION TRENDS OF 2 OR MORE')
      WRITE(5,6)
6     FORMAT(1X,'OCCUPATIONS TO DETERMINE WHETHER THEY CAN BE'
     1  ' GROUPED TOGETHER FOR LOG-')
      WRITE(5,7)
7     FORMAT(1X,'PROBABILITY ANALYSIS.')
      WRITE(5,10)
10    FORMAT(/,1X,'FOR PURPOSES OF THIS PROGRAM:',/,6X,'(1) '
     1  'THE "X - VALUES" = LENGTH OF SERVICE COMPLETED'
     2  '; AND')
      WRITE(5,11)
11    FORMAT(6X,'(2) THE "Y - VALUES" = NUMBER (OR PERCENT)'
     1  ' RETAINED AT TIME X.')
      WRITE(5,12)
12    FORMAT(/,1X,'SEE INSTRUCTION MANUAL FOR FURTHER EXPLANA'
     1  'TION OF DATA REQUIRED.')
20    WRITE(5,30)
30    FORMAT(///,1X,'ENTER THE NO. OF RETENTION GROUPS TO BE COMPARED')
      READ(5,50)NC
50    FORMAT(I2)
      WRITE(5,60)
60    FORMAT(//,1X,'ENTER THE OCCUPATION CODE FOR GROUP')
      DO 65 I=1,NC
      WRITE(5,140)I
      READ(5,62)IOC(I)
62    FORMAT(I5)
65    CONTINUE
      WRITE(5,70)
70    FORMAT(/,1X,'ENTER THE NO. OF X-VALUES')
      READ(5,90)NX
90    FORMAT(I2)
      WRITE(5,110)
110   FORMAT(//,1X,'ENTER X-VALUE')
      DO 180 I=1,NX
      WRITE(5,140)I
140   FORMAT(/,1X,'NO. ',I2,':')
      READ(5,160)X(I)
160   FORMAT(F7.4)
      XLOG(I)=ALOG10(X(I))
180   CONTINUE
      WRITE(5,200)
200   FORMAT(//,1X,'ENTER STARTING POPULATION (N) FOR GROUP')
      DO 260 I=1,NC
      WRITE(5,230)I,IOC(I)
230   FORMAT(/,1X,'NO. ',I2,':  (',I4,')')
      READ(5,250)NO(I)
250   FORMAT(I6)
260   CONTINUE
      WRITE(5,280)
```

```
280     FORMAT(//,1X,'ARE Y-VALUES IN',/,4X,'(1) NUMBER OR',/,
       1  4X,'(2) PERCENT FORM?',/,4X,'(ANS 1 OR 2)')
        READ(5,300)IP
300     FORMAT(I2)
        IF(IP.EQ.2)GO TO 460
        WRITE(5,330)
330     FORMAT(//,1X,'IN INTEGER FORM (NO DEC. PTS.)',/,1X,
       1  'AND SEPARATED BY COMMAS, ENTER THE',/,1X,
       2  'Y-VALUES CORRESPONDING TO X =')
        DO 440 I=1,NX
        WRITE(5,360)X(I)
360     FORMAT(/,1X,F7.4,':')
        READ(5,*)(IY(I,J),J=1,NC)
        DO 440 K=1,NC
        EN=NO(K)
        Y(I,K)=IY(I,K)
        Y(I,K)=Y(I,K)/EN
        YSD(I,K)=ANDPX(Y(I,K))
440     CONTINUE
        GO TO 580
460     WRITE(5,470)
470     FORMAT(//,1X,'IN DECIMAL FORM, SEPARATED BY COMMAS',/,
       1  1X,'ENTER THE Y-VALUES CORRESPONDING TO X =')
        DO 570 I=1,NX
        WRITE(5,500)X(I)
500     FORMAT(/,1X,F7.4,':')
        READ(5,*)(Y(I,J),J=1,NC)
        DO 570 K=1,NC
        EN=NO(K)
        YSD(I,K)=ANDPX(Y(I,K))
570     CONTINUE
580     WRITE(5,585)
585     FORMAT(//,1X,'DO YOU WISH TO TEST (1) ALL OR (2) '
       1  'SOME OF THESE GROUPS?',/,1X,'(ANS 1 OR 2)')
        READ(5,300)MS
        IF(MS.EQ.2)GO TO 595
        DO 590 IQ=1,NC
590     NUM(IQ)=IQ
        NG=NC
        GO TO 610
595     WRITE(5,600)
600     FORMAT(//,1X,'HOW MANY GROUPS DO YOU WISH TO TEST?')
        READ(5,300)NG
        WRITE(5,605)
605     FORMAT(//,1X,'WHICH ONES?  ENTER USING THE GROUP NUM'
       1  'BERS',/,1X,'ESTABLISHED DURING THE ENTRY OF'
       2  ' OCCUPATION CODES.')
        WRITE(5,607)
607     FORMAT(1X,'(SEPARATE THESE VALUES WITH COMMAS.)')
        READ(5,*)(NUM(I),I=1,NG)
610     ENTOT=0.0
        NSUM=NC+1
        DO 630 I=1,NG
        LR=NUM(I)
        ENO=NO(LR)
```

```
630       ENTOT=ENTOT+ENO
          NO(NSUM)=ENTOT
          DO 710 I=1,NX
          Y(I,NSUM)=0.0
          DO 680 J=1,NG
          LQ=NUM(J)
          ENO=NO(LQ)
          YD(I,LQ)=Y(I,LQ)*ENO
          Y(I,NSUM)=Y(I,NSUM)+YD(I,LQ)
680       CONTINUE
          Y(I,NSUM)=Y(I,NSUM)/ENTOT
          YSD(I,NSUM)=ANDPX(Y(I,NSUM))
710       CONTINUE
          KS=1
720       LS=NUM(KS)
730       SUMX=0.0
          SUMY=0.0
          SUMXY=0.0
          SUMX2=0.0
          DO 820 I=1,NX
          SUMX=SUMX+XLOG(I)
          SUMY=SUMY+YSD(I,LS)
          SUMXY=SUMXY+XLOG(I)*YSD(I,LS)
          SUMX2=SUMX2+XLOG(I)**2
820       CONTINUE
          ENI=NX
          S1=ENI*SUMX2-SUMX*SUMX
          S3=ENI*SUMXY-SUMX*SUMY
          B(KS)=S3/S1
          A(KS)=SUMY/ENI-(B(KS)*SUMX)/ENI
          IF(KS.EQ.NSUM)GO TO 885
          KS=KS+1
          IF(KS.LE.NG)GO TO 720
          KS=NSUM
          LS=NSUM
          GO TO 730
885       JP=0
          DO 890 MM=1,NX
          DO 890 IK=1,NG
          LL=NUM(IK)
          YPT(MM,LL)=ANDXP(YSD(MM,LL))
890       YPT(MM,LL)=YPT(MM,LL)*100.
          DO 895 IG=1,NG
          JL=NUM(IG)
          SMP(JL)=0.0
          SMP2(JL)=0.0
          DO 895 KL=1,NX
          SMP(JL)=SMP(JL)+YPT(KL,JL)
          SMP2(JL)=SMP2(JL)+YPT(KL,JL)**2
895       CONTINUE
          SUMP=0.0
          SUMP1=0.0
          SUMP2=0.0
          DO 900 KA=1,NG
          KB=NUM(KA)
```

```
LPTEST
         SUMP=SUMP+SMP(KB)
         SUMP2=SUMP2+SMP2(KB)
 900     SUMP1=SUMP1+SMP(KB)**2
         PC2=SUMP**2
         EN1=NX*NG
         ENK=NX
         EN2=NG*(NX-1)
         EN3=NG-1
         VAM=SUMP2-SUMP1/ENK
         VAB=SUMP1/ENK-PC2/EN1
         VARM=VAM/EN2
         VARB=VAB/EN3
         F2=VARB/VARM
         NGMS=NX
         NBAS=NX*NG
         IF(F2.LE.1.)GO TO 910
         CALL FTEST(NGMS,NBAS,F2,EP)
         IF(EP.LE.0.05)JP=1
         GO TO 999
 910     EP=1.
 999     WRITE(5,1000)(NXX(I),I=1,21)
 1000    FORMAT(/////,1X,21A2)
         WRITE(5,1010)
 1010    FORMAT(1X,´X´,40X,´X´)
         WRITE(5,1010)
         WRITE(5,1020)
 1020    FORMAT(1X,´X  -------LPTEST ANALYSIS OUTPUT-------´
         1  ´  X´)
         WRITE(5,1010)
         WRITE(5,1010)
         WRITE(5,1030)NG
 1030    FORMAT(1X,´X  THE ´,I2,´ OCCUPATIONS TESTED´,13X,´X´)
         DO 1050 I=1,NG
         LU=NUM(I)
         WRITE(5,1040)IOC(LU)
 1040    FORMAT(1X,´X´,6X,I5,29X,´X´)
 1050    CONTINUE
         IF(JP.EQ.0)WRITE(5,1060)
 1060    FORMAT(1X,´X  MAY BE GROUPED TOGETHER.´,14X,´X´)
         IF(JP.EQ.1)WRITE(5,1070)
 1070    FORMAT(1X,´X  MAY NOT BE GROUPED TOGETHER.´,10X,´X´)
         WRITE(5,1010)
         WRITE(5,1010)
         WRITE(5,1080)
 1080    FORMAT(1X,´X  INDIVIDUAL SUBGROUP DATA:´,13X,´X´)
         WRITE(5,1010)
         WRITE(5,1090)
 1090    FORMAT(1X,´X´,20X,´--L - P EQUATION--´,2X,´X´)
         WRITE(5,1100)
 1100    FORMAT(1X,´X´,6X,´OCCN´,4X,´N´,8X,´(A)´,7X,´(B)´,
         1  4X,´X´)
         WRITE(5,1110)
 1110    FORMAT(1X,´X´,6X,´----´,4X,´-´,8X,´---´,7X,´---´,
         1  4X,´X´)
         WRITE(5,1010)
```

```
        DO 1130 M=1,NG
        JM=NUM(M)
        JOC=IOC(JM)
        JN=NO(JM)
        WRITE(5,1120)JOC,JN,A(M),B(M)
1120    FORMAT(1X,'X',5X,I5,2X,I5,3X,F8.5,2X,F8.5,2X,'X')
1130    CONTINUE
        IF(JP.EQ.1)GO TO 1155
        WRITE(5,1010)
        WRITE(5,1010)
        WRITE(5,1140)
1140    FORMAT(1X,'X  TOTAL GROUP EQUATION:',17X,'X')
        WRITE(5,1010)
        B(NSUM)=ABS(B(NSUM))
        WRITE(5,1150)A(NSUM),B(NSUM)
1150    FORMAT(1X,'X',6X,'Y = ',F8.5,' - ',F8.5,'X',10X,'X')
1155    WRITE(5,1010)
        WRITE(5,1010)
1160    WRITE(5,1165)(NXX(I),I=1,21)
1165    FORMAT(1X,21A2)
        WRITE(5,1170)
1170    FORMAT(////////,1X,'DO YOU WISH TO TEST ANOTHER '
        1  'SUBSET OF THESE GROUPS? (Y OR N)')
        READ(5,1180)LAG
1180    FORMAT(A2)
        IF(LAG.EQ.'Y')GO TO 580
        WRITE(5,1190)
1190    FORMAT(//,1X.'DO YOU WISH TO TEST ANOTHER SET OF'
        1  ' GROUPS? (Y OR N)')
        READ(5,1180)NAG
        IF(NAG.EQ.'Y')GO TO 20
        STOP
        END
```

```
      SUBROUTINE FTEST(NGMS,NBAS,F,P)
C     APPROXIMATES THE PROBABILITY OF CHANCE OCCURRENCE OF THE
C     OBSERVED VALUE OF F, GIVEN NGMS (= N OF GROUP TESTED -- BY
C     DEFINITION, THE GREATER MEAN SQUARE) AND NBAS (= N OF THE
C     BASE GROUP).
C     REQUIRES SUBPROGRAM: FUNCTION ANDXP.
C     (REFERENCE: NBS, ABRAMOWITZ & STEGUN,AMS 55,9TH,26.6.15,P.947.)
      V1=NGMS-1
      V2=NBAS-1
      EX=1./3.
      F3=F**(EX)
      F23=F3**2
      B1=2./(9.*V1)
      B2=2./(9.*V2)
      T1=1.-B2
      T2=1.-B1
      TOP=F3*T1-T2
      BOT=SQRT(B1+F23*B2)
      X=TOP/BOT
      Q=ANDXP(X)
      P=1.-Q
      RETURN
      END
      FUNCTION ANDPX(P)
C     ACCUMULATIVE NORMAL DISTRIBUTION FUNCTION.
C     APPROXIMATES X (+ OR - .000004) FROM P. (HASTINGS, P192)
      DATA A0,A1,A2,B1,B2,B3/2.515517,.802853,.010328,1.432788,
     1 .189269,.001308/
      IF(P.GE.1.) X1=4.
      IF(P.GE.1.) GO TO 380
      IF(P.GT.0.0005) GO TO 70
      X1=-4.
      GO TO 380
70    IF (P.LT.0.5) GO TO 110
      E = SQRT(ALOG (1./(1.-P)**2))
      X1 = E-((A2*E+A1)*E+A0)/(((B3*E+B2)*E+B1)*E+1.)
      GO TO 140
110   P = 1.-P
      E = SQRT(ALOG(1./(1.-P)**2))
      X1 = -1.*(E-((A2*E+A1)*E+A0)/(((B3*E+B2)*E+B1)*E+1.))
      P = 1.-P
140   AX1 = X1 +0.0005
      AX2 = X1 -0.0005
      AP1 = ANDXP(AX1)
      AP2 = ANDXP(AX2)
      AI = (P-AP2)/(AP1-AP2)
      BX1 = AX2+AI*(AX1-AX2)
      D1 = P-ANDXP(BX1)
      IF (D1.LT.0.000001) GO TO 290
      AX3 = BX1+1.1*D1
      AX4 = BX1-1.1*D1
      AP3 = ANDXP(AX3)
      AP4 = ANDXP(AX4)
      AII = (P-AP4)/(AP3-AP4)
      BX1 = AX4+AII*(AX3-AX4)
      D1 = P-ANDXP(BX1)
290   IF (D1.LE.0.0000005) GO TO 370
      IF (D1.GT.0.0000005) GO TO 340
      BX1 = BX1+0.0000002
      D1 = ANDXP(BX1)
      GO TO 290
340   BX1 = BX1-0.0000002
      D1 = ANDXP(BX1)
      GO TO 290
370   ANDPX = BX1
      GO TO 390
380   ANDPX=X1
390   RETURN
      END
      FUNCTION ANDXP(X)
C     ACCUMULATIVE NORMAL DISTRIBUTION FUNCTION. APPROXIMATES P
C     FROM X (TO + OR - .0000001). (NBS-55, P. 932)
      X1 = ABS(X)
      T = 1./(1.+.2316419*X1)
      ANDXP = 1.-.3989423*EXP(-(X1**2)/2.)*(.3193815*T-.3565638*T**2
     1  +1.781478*T**3-1.821256*T**4+1.330274*T**5)
      IF (X.LT.0.0) ANDXP = 1.-ANDXP
      RETURN
      END
```

APPENDIX F

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

GS810

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

GS810

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

GS810 TECHNICAL ANALYSIS

Purpose

Computer terminal program GS810 is designed to enable personnel workers without extensive statistical background to:

(1) Make detailed quantitative analyses and projections of the grade-advancement patterns of individual occupations or employee groups for use in recruitment and counselling programs, workforce and budget studies, and organization staffing needs planning programs;

(2) Make objective comparisons of the grade/time pattern of one occupation or employee group with that of another, for use in equal employment opportunity, career planning and in-service placement programs; and

(3) Identify by objective means, for purposes of executive development, performance evaluation, and occupational standards study programs, those individuals or subgroups within an occupation or organization whose advancement trends have been signficantly above or signficantly below the prevailing norm for this group.

In accordance with these purposes, GS810 has been designed (with its companion subprograms) as a comprehensive and self-contained unit which will do the complete job of turning raw input data into the finished output projections and evaluations desired by the program operator without any need for operator technical participation and without any need for operator reference to outside assistance sources (statistics texts, look-up tables, etc.) in order to evaluate the significance of calculation results.

Method

Input Files - Data are input to the program from previously stored 8 x 10 integer files which are accessed by means of a 6-space file-name code (one code system is explained in the operation manual). By this means, the operator need do no data entry: only type in the code name of the (previously-entered and -verified) files chosen for study and they are automatically read into the program.

This technique is extremely flexible since it frees the operator to select and combine any files (up to a maximum of 8) and in any order or combination. The technique is also extremely reliable because all data items in each file are brought into the program with 100% accuracy; i.e., without the inevitable errors of manual entry. And, of course, this technique is far more efficient in the accessing and manipulation of files than is the case with laborious manual data-entry methods.

Files are in integer format for maximum ease and accuracy of data entry and for minimum-space storage. Since people-count tables, characteristically, involve no decimal fractions, integer format (no decimal points specified) is (a) sufficient, (b) saves both the work and the inevitable errors of entering decimal points, and (c) conserves one digit of storage space for every number entered.

Data Arrangement - The program is formatted for 8-grade by 10-year grade distribution tables ("8 x 10 arrays" in computer usage). An 8-grade pattern was selected to provide full coverage of the eight grade-advancement steps from grade 5 to 15* in the General Schedule pattern. Any other uniform-step, equal-interval progression pattern desired, however, whether involving grades or, in the alternative, salary categories, can be entered by the operator. If less than 8 grades (rows) of data are to be used, the unused rows in the files should simply be filled with zeroes (0,0,0...etc.)

A "first-10-year" pattern was selected for the length-of-service (columnar) dimension for several reasons:

(1) The most rapid grade rise in any group occurs in the first 10 years of service; thus the rise of the group's characteristic advancement curve is largest and, hence, most accurately measured in this initial period.

(2) This span covers those length-of-service groups which are the numerically largest in most organizations: groups which are of special interest in both analysis and projection because of their importance for the organization's expected future. And

(3) The span of the last ten years encompasses most or all of the period of our most intensive national efforts toward equal employment opportunity: data for this period thus constitute the clearest available test of both (a) the objective results of recent hiring and promotion policies, and (b) the probable results to be expected from these policies for the future.

As with unused r[illegible], as mentioned above, if less than 10 years' data are available (at least two are a must), the unused file columns should simply be filled out with zeroes. Similarly, also, if column-values other than whole - or half-years are desired (e.g., where there are irregular time intervals between columns), the desired column-values can be entered by the operator.

---

*So that grade advancement curves in the "two-grade-interval" occupations follow the same "one digit-one step" rate below GS-11 that they do above GS-11 (which is essential for curve computation purposes) the two-digit-interval grades of GS-5,7, and 9 are converted to grade weights of "8", "9", and "10" respectively.

Trend Analysis - Since grade-advancement trends are rising curves whose rate of increase slows rapidly over time, the appropriate curve form for such trends is one of the family of "growth" or "maturation" curves. The two most widely-used in economics and population studies are the Pearl-Reed Logistic and the Gompertz:

(1) $1/Y = a+bc^x$ (Logistic)
(2) $\log Y = \log a + \log b(c^x)$ (Gompertz)

The specially useful feature of these familiar curves, of course, is the inclusion, among the multiplicative terms, of the additive term "a." From this term the curve asymptote--the maximum ultimate value toward which the curve is tending as the increasing value of x makes c^x approach 0*--is easily determined. It is this additive character of this term, however, which makes it possible to fit such curves to actual data only by approximate methods, and even then, only

(a) when the total number of observations is some multiple of three (6, 9, 12, etc),

(b) when all time intervals between observations are exactly equal, and

(c) when no observations in the sequence are missing.

In purely practical terms, then, such severely restrictive terms are, to say the least, not always met in the average personnel management operating situation. The use of these commonly-seen curves is therefore inappropriate for a program intended for a wide range of applications.

(As a final consideration, it is technically sufficient to note that the deviation of the observed values from the trend lines fitted by the approximate methods necessary for these curves

(a) are not minimal in size, as in least-squares fits, and

(b) are not equal in sum on both sides of the trend line, as needed for reliable tests of variance.)

A third type of maturation curve is a relative of the log-probability decay curve which we have established** as the exact form of the work group retention curve:

---

*The c term is always less than 1, thus c^x rapidly becomes smaller with increasing values of x.

**H.L. Clark, "Problems and Progress in Civil Service Manpower Planning," NATO Conference on Manpower Planning Models, Cambridge, U.K., Sept. 1971.

$$(3)\ z(y) = a + b \log x$$

However, this curve can be fitted only to longitudinal data (where all data are observations of the same group at different points in time) and only when both the original size of the group and its size at the time of each grade observation are known.

The fourth growth curve form is the general category of the exponential, which includes such forms as

$$(4)\ Y = ab^{x} \text{ or } Y = ax^{b}$$

These curves can be converted, with the use of logarithms, to the linear forms:

$$(4.a)\ \log Y = \log a + X \log b \text{ or}$$
$$(4.b)\ \log Y = \log a + b \log X$$

Since such linear forms do permit the use of least-squares fitting techniques, they can readily be applied:

(1) To cross-section data, as well as longitudinal data, where the observed population in one column typically may differ considerably from the populations of other columns,

(2) Where observations are separated by unequal time intervals, and

(3) Where one or more observations in a series are not available (a missing year, etc.)

Because of this linear form, the extrapolation of such curves is technically very easy. And because they are fitted by least-squares techniques, variance analysis is straightforward.

As a final note, the closeness with which curves of this equation form can be fitted to the 10-year data used in this program is statistically scarcely distinguishable, for moderate to small size samples, from the fits obtainable from the log-probability curve form.

Computations

GS810 performs a series of computations on the data input from the called files:

(1) The files are added together to produce a total grade distribution table of all entered groups, with the row-values (i.e., grades or weights) and the column-values (i.e., years) specified by the operator.

(2) The row-values (Y-values) and the column-values (X-values) are then converted to logs and a weighted linear trend line of form 4.b is calculated. (The coefficient of linear correlation (r) is calculated as a measure of the proportion of total variance which is accounted for by the trend line.)

(3) The value of the regression equation is then calculated for each column. The squared deviation of each block of column data from the regression line is determined, multiplied by the block frequency, and the overall sum of variances is divided by the total group N-1 to produce the overall group variance. The program then prints out (1) the equation, (2) r, (3) N, and (4) the group variance.

(4) At operator option, the program iterates regression line over 35 years.

(5) At operator option, the program fits the overall group "norm" line of regression to the distribution of each subgroup and determines the variance of the subgroup from the overall "norm." This subgroup variance is then compared with the variance of the overall group from the same curve and the significance of the difference is measured.

Significance Test - The significance of the ratio of the individual subgroup variance to the total group variance is evaluated by means of the F-ratio test, one of the most fundamental, reliable and flexible of statistical tests.

The value of F is given by:

$$F = \frac{\text{Variance of subgroup}}{\text{Variance of total group}}$$

with n_1 (Subgroup degrees of freedom) = $N_s - 2$

n_2 (Total group degrees of freedom) = $N_t - 2$

The probability - value of the resulting F-ratio is calculated by means of the Q(F) approximation given by Abramowitz and Stegun (26.6.15 AMS 55, 9th, 1970) and the significance of Q is evaluated with the usual fiduciary limit of $p = .05$. (The variance of the total group is used as the denominator in order to minimize the probability of "Beta" error: the apparent finding of a "difference" where in fact no difference exists.)

Sample Size - The use of the F-test to evaluate the results of the program's calculations is one of GS810's key features, not only because of this test's well-known power and validity, but--most especially--because of its ability to produce valid results from samples smaller than those of almost any other statistical test : once the overall group "norm" advancement curve is known, reliable comparisons with this curve can be made by means of the F-test for subgroup samples as small as 3.

Special Program Features

Certain special features of GS810, in addition to its overall design as a completely self-contained system, are worth special mention:

1. The operation manual describes a file code system the operator may use in naming, storing and calling files.
2. The program includes detailed step-by-step guidance to the operator in the selection of column- and row-values.
3. The program offers the operator a wide range of output and analysis options, each of which will be executed automatically at the operator's choice:
    - a.) A write-out of the complete grade distribution of the overall group (i.e., the sum of all entered files).
    - b.) A 35-year career-progression projection of the computed "norm" curve.
    - c.) An F-ratio test of each subgroup relative (1) to the "norm" curve pattern of the entire group or (2) to the entered equation and variance of an outside group.
    - d.) A write-out of the complete grade distribution of each individual group entered into the program
4. Every output or "results" section written out by the program includes the full file-code name of the group to which it refers (including, if the group is the sum of subgroups, the full file-code name of every subgroup included the total group).

Automatic Run Option

When each program run-through with a group of files is completed, the operator who asks for another run with new files is offered the option of having that next run in an automatic mode: once he specifies the new files he wishes to analyze and the code he wishes to use for the sum of the entered subgroups, the entire program will thereafter run through to completion, with the exact same pattern of analysis and outputs which he specified in his previous step-by-step run-through, on a completely automatic basis without further operator participation.

On each run-through of GS810, that is, the program records every program option chosen by the operator and if no change is desired in this pattern, each subsequent group of files called by the operator will be processed in exactly the same way completely automatically. With this option, then, any desired pattern of options can be repeated with successive groups and combinations of files with an absolute minimum of operator effort and with a maximum of speed and efficiency.

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

GS810

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

DATA ARE ENTERED INTO THE "GS810" PROGRAM FROM PREVIOUSLY-STORED FILES. THESE FILES CONSIST OF 8 ROWS AND 10 COLUMNS. EACH ROW REPRESENTS A GRADE-LEVEL; EACH COLUMN A SPECIFIC LENGTH-OF-SERVICE CATEGORY.

FOR USE WITH THE "GS810" PROGRAM, DATA MUST BE COLLECTED BY LENGTH-OF-SERVICE CATEGORIES. A COUNT SHOULD BE MADE OF THE NUMBER OF EMPLOYEES IN EACH SUCH CATEGORY (E.G., 0-1 YEARS, 1-2 YEARS, ETC.) FOR EACH GRADE-LEVEL (E.G., GS-5,7,9, ETC.). SUCH COUNTS SHOULD BE MADE FOR EACH OF THE MAJOR OCCUPATIONS IN AN ORGANIZATION. (THESE DATA MAY COME FROM EXISTING AUTOMATED SYSTEMS OR FROM ANY OTHER SOURCE AVAILABLE TO AN ORGANIZATION.)

A SAMPLE OF SUCH A COUNT MIGHT BE:

OCCUPATION : GS-00XXX

	LENGTH OF SERVICE (YEARS)									
GRADE	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10
5	3	2								
7	4	2	1							
9	5	3	2	1	1					
11	6	4	1		1	2				
12	6	3	1	2		1				
13		2	1							
14			2		1					
15								2		

THIS TABLE CAN VERY SIMPLY BE CONVERTED TO A "GS810"-TYPE FILE. SUCH FILES CONSIST OF 8 CONSECUTIVE TYPED LINES (ONE / GRADE-LEVEL) OF 10 ENTRIES EACH (ONE / LENGTH-OF-SERVICE CATEGORY), SEPARATED BY COMMAS, WITH EACH ENTRY MADE UP OF 1-4 INTEGERS. ZEROES ARE USED WHERE BLANKS OCCUR IN THE TABLE.

A COMPUTER FILE FOR THE ABOVE TABLE WOULD BE:

```
3,2,0,0,0,0,0,0,0,0,
4,2,1,0,0,0,0,0,0,0,
5,3,2,1,1,0,0,0,0,0,
6,4,1,0,1,2,0,0,0,0,
6,3,1,2,0,1,0,0,0,0,
0,2,1,0,0,0,0,0,0,0,
0,0,2,0,1,0,0,0,0,0,
0,0,0,0,0,0,0,2,0,0
```

EXECUTION COMMANDS:

TO BEGIN EXECUTION OF THE "GS810" PROGRAMMING SEQUENCE, AN OPERATOR WILL PERFORM A CHAIN OF EXECUTION COMMANDS. THE ACTUAL FORM OF THESE COMMANDS WILL DEPEND ON THE TIME-SHARING SYSTEM BEING USED. IN GENERAL, THESE COMMANDS WILL PERFORM THE FOLLOWING OPERATIONS:

- CALL UP "GS810" AND ITS FOUR SUBPROGRAMS "RLL810", "VLL810", "FTEST" AND "ANDXP" AND TRANSLATE THEM INTO MACHINE LANGUAGE. THISIS THE COMPILATION PHASE.
- LOAD THE COMPILED PROGRAM AND SUBPROGRAMS INTO THE CENTRAL PROCESSING AREA AND START PROGRAM RUN. THIS IS THE EXECUTION PHASE.

DATA ENTRY:

*THIS PROGRAM ANALYZES GRADE/ *TIME TRENDS IN 1-8 GROUPS OF 3 OR *MORE EACH; USES 8-GRADE X 10- *YEAR DATA ARRAYS.	SELF-EXPLANATORY

*ENTER FILES VIA 6-SPACE CODE (SEE INSTR).

THE SUGGESTED FILE CODE CONSISTS OF:

SP (1)= (M)INORITY, (N)ON-MIN, (N)O., ETC.
SP (2)= (M)ALE, (F)EMALE, N(O)., (T)OTAL, ETC.
SP (3)-(5)= GS SERIES CODE: E.G., (2)(0)(1). (1810=810)
SP (6)= YEAR: 7(4),7(5),OR (0) (=74+75)

ABOVE SELF-EXPLANATORY. USER MAY CHOOSE TO USE DIFFERENT FILE CODE SYSTEM. IF SO, CODE NAMES MUST CONSIST OF NO MORE THAN 6 LETTERS AND/OR DIGITS, IN ANY DESIRED COMBINATION.

*HOW MANY FILES DO YOU WISH TO *ANALYZE? (ANS 1-8)	SELF-EXPLANATORY.
*PLEASE ENTER CODE NAME OF EACH *FILE: *1=	ENTER 6-SPACE CODE NAME OF FIRST DESIRED FILE. FILES WILL BE NUMBERED IN ORDER OF ENTRY.
*(ERROR MESSAGE)	FILE-READ ERROR; RUN ABORTED. CHECK FOR FILE-CODE TYPING ERROR (ESP: O FOR 0), WRONG CODE, OR BAD FILE. TO CHECK FILE, TYPE OUT THE FILE AND CHECK FOR 8 LINES, 10 ENTRIES, MISSING NO.'S OR COMMAS, OR NO.'S WITH MORE THAN 4 DIGITS.

* = COMPUTER WRITTEN

```
*DATA TAKEN FROM FILES. NEXT:
*
*"STD PATTERN" DESIRED? (Y OR N)
```

IF THE 'Y' OPTION IS CHOSEN, THE COMPUTER ASKS:

```
*(1) 5/15 OR (2) 1/8? (1 OR 2)
```

THE "STD PATTERN" OPTION ALLOWS THE OPERATOR TO SET AT THE BEGINNING OF THE PROGRAM CERTAIN BASIC DECISION VARIABLES TO THE VALUES THEY MOST FREQUENTLY HAVE. THIS OPTION PROVIDES A FASTER RUN THROUGH OF THE PROGRAM. (FURTHER EXPLANATION OF THE VARIABLES INVOLVED IN THIS OPTION IS FOUND ON PAGE 6.)

IF 'N' IS CHOSEN, THE COMPUTER ASKS:

```
*    (A) SELECT COLUMN-VALUE PATTERN WANTED:
*        (USE COL. MIDPOINTS: "0-1"YRS= "0.5", ETC.)
*
*        (1) 0.5,1.5, ETC.
*        (2) 1.,2., ETC.
*        (3) OTHER (1,2 OR 3)
```

MOST APPLICATIONS OF THIS PROGRAM WILL INVOLVE THE DIVIDING OF EMPLOYEES INTO LENGTH-OF-SERVICE GROUPINGS (COLUMNS). SINCE SUCH GROUPINGS COME FROM A CONTINUOUS DISTRIBUTION, EACH COLUMN WILL REPRESENT A SPECIFIED RANGE OF VALUES: 0.0-0.99 YEARS OF SERVICE, 1.00-1.99 YEARS, AND SO ON. FOR COMPUTATION PURPOSES, THAT ONE VALUE WITHIN A GIVEN COLUMN RANGE WHICH BEST REPRESENTS THE AVERAGE OF THOSE INCLUDED IN THE COLUMN IS ASSUMED TO BE THE RANGE MIDPOINT: 0.5 FOR A 0-1 RANGE, 1.0 FOR A 0.5-1.5 RANGE, AND SO FORTH. IT IS THIS MIDPOINT WHICH SHOULD BE USED AS THE COLUMN VALUE HERE.

IN SOME CASES, THE "0.5/1.5" OR THE "1./2." PATTERNS MAY NOT BE APPROPRIATE: E.G., WHERE COLUMNS COVER PERIODS OTHER THAN YEARS OR WHERE THE INDIVIDUAL COLUMNS REPRESENT SAMPLES TAKEN AT IRREGULARLY-SPACED POINTS IN TIME. IN SUCH CASES, OPTION "3" SHOULD BE USED. WHEN "3" IS ENTERED, COMPUTER REPLIES:

```
*     ENTER DESIRED VALUE FOR MIDPOINT OF COL:
*
*     1
*     2
      ETC.
```

NOTE THAT DECIMAL POINTS MUST -- REPEAT MUST -- BE SHOWN FOR ALL ENTRIES IN THIS OPTION: THEIR OMISSION WILL RESULT IN COMPUTATION ERRORS.

FILES MAY CONTAIN FEWER THAN 10 COLUMNS OF DATA WITHOUT HARMING COMPUTATION PROGRAM OR STATISTICAL TESTS; SIMPLY FILL ANY UNUSED COLUMNS WITH 0'S (ZEROES).

* = COMPUTER WRITTEN

THE "3" OPTION MAY ALSO BE USED TO ENTER COLUMN-VALUES FOR PERIODS OTHER THAN THE FIRST 10 YEARS OF SERVICE (11-20 YEARS, 5-15, ETC.).

NEXT :

```
*   (8) SELECT GRADE PROGRESSION PATTERN
*       (ROW-VALUE) PATTERN WANTED:
*
*       (1) GS-5/15
*       (2) GS-1/8
*       (3) OTHER (1,2 OR 3)
*
```

SELF-EXPLANATORY. IF THE OPERATOR ANSWERS "1", COMPUTER NOTES:

```
*   PATTERN APPLIED:
*
*   SO THAT "2-GRADE-INTERVAL-SERIES" GRADES PROGRESS AT THE SAME
*   "1-GRADE/1-STEP" RATE BELOW GS-11 AS THEY DO ABOVE GS-11,GS-5/9
*   GRADES WERE WEIGHTED: GS-5="8", GS-7="9", AND GS-9="10".
*   (0.50 WAS ADDED TO EACH GRADE-VALUE AS THE MIDPOINT OF THE
*   TOTAL RANGE OF JOBS IN THAT GRADE.)
```

THE TREND OF GRADES OVER TIME CAN BE QUANTITATIVELY ANALYZED WITH FULL EFFECTIVENESS ONLY WHEN THE NUMERICAL WEIGHTS OF THE GRADES ARE PROPORTIONAL TO THE ACTUAL NUMBER OF STEPS INVOLVED IN THE GRADE-ADVANCEMENT LADDER. SINCE THE "TWO-GRADE-INTERVAL" OCCUPATIONS BY DEFINITION PROGRESS BY STEPS GS-5, 7, 9, 11, 12, 13, ETC., THE REPLACEMENT OF GRADES BY WEIGHTS IN THE GS-5 TO GS-9 RANGE IS ESSENTIAL TO AVOID NUMERICAL MISREPRESENTATION OF THE ACTUAL STRUCTURE OF THE CAREER LADDER.

IF THE OPERATOR ANSWERS "2" TO THIS QUESTION, THE NUMERICAL WEIGHTS USED REFLECT THE ACTUAL NUMBER OF STEPS INVOLVED IN THE GRADE ADVANCEMENT LADDER FOR THE LOWER-GRADED OCCUPATIONS. IN THIS CASE, THE GRADES 1, 2 AND 5 - 10 REPRESENT HALF STEPS WHILE GRADES 3 AND 4 REPRESENT FULL STEPS. THESE GRADES ARE WEIGHTED ACCORDINGLY.

IN OPTION "3", THE OPERATOR ENTERS THE EXACT WEIGHTS WANTED ONE BY ONE:

```
*   ENTER MIDPOINT (E.G.,GRD."1"=
*   (1.-1.99)="1.5") FOR ROW:
*   1=
*   2=
     (ETC.)
```

AS WITH COLUMN-VALUES, DECIMAL POINTS MUST BE SHOWN WITH ALL ENTRIES OR COMPUTATION ERRORS WILL RESULT. THE HIGHEST NUMBER SHOULD NOT EXCEED 99.99.

* = COMPUTER WRITTEN

```
*TEST SUBGROUPS AGAINST:
*     (1) THEIR OWN SUM?
*     (2) AN OUTSIDE GROUP?
*     (3) NO TESTS DESIRED.  (1,2 OR 3)
```

OPTION 2 ABOVE REQUIRES THAT "OUTSIDE GROUP" BE CLEARLY DEFINED:

"OUTSIDE GROUP" = A GROUP WHICH (A) IS NOT INCLUDED AMONG THE GROUPS WHOSE FILES ARE ENTERED IN THIS PROGRAM, BUT (B) WHOSE "VARIANCE FROM THE FITTED LINE OF REGRESSION" HAS BEEN COMPUTED FROM DATA WHICH INCLUDED ALL OF THE DATA -- REPEAT, ALL: WITHOUT ANY EXCEPTION WHATSOEVER -- WHICH IS CONTAINED IN THE FILES ENTERED IN THIS PROGRAM.

STRICT OBSERVANCE OF THE LETTER OF THIS REQUIREMENT IS ESSENTIAL TO GET VALID RESULTS. THE EXTREMELY WIDE, ALMOST UNLIMITED RANGE OF APPLICATION OF THIS PROGRAM IS IN LARGE PART DUE TO THE EXTREMELY WIDE RANGE OF SAMPLE SIZES (DOWN TO SAMPLES AS SMALL AS 3) OVER WHICH THE VARIANCE-RATIO ("F" STATISTIC) CAN BE VALIDLY TESTED. THE USE OF THE F-TEST WITH VARIANCES COMPUTED FROM LINES OF REGRESSION IS ONLY VALID, HOWEVER, WHEN THE VARIANCE OF THE BASE (THE "OUTSIDE" GROUP) IS COMPUTED FROM A CURVE FITTED TO DATA WHICH INCLUDED ALL OF THE DATA OF EACH SUBGROUP (I.E., EACH SAMPLE) WHICH IS TO BE TESTED HERE. IF THESE CONDITIONS ARE ALL FULLY MET, AN ANSWER OF "2" WILL YIELD THIS SEQUENCE:

```
*WHAT 6-SPACE CODE NAME SHOULD WE USE FOR THE
*OUTSIDE (BASE) GROUP?
*
*PLEASE ENTER THE FOLLOWING FOR BASE GROUP (NAME) :
*
*     (NAME) GRADE/TIME EQUATION
*     (E.G., 1.12345,.123456):
*
*     (NAME) N, VARIANCE (E.G.: 123,1.12345):
```

IN ALL CASES WHERE IT IS NOT ABSOLUTELY CLEAR THAT THE CONDITIONS FOR OPTION 2 ARE FULLY MET, AS WELL AS IN ALL CASES WHERE GROUP-SUM COMPARISONS ARE DESIRED, OPTION 1 SHOULD BE SELECTED. IN THIS OPTION, THE PROGRAM WILL ADD ALL OF THE ENTERED FILES INTO AN OVERALL TOTAL DISTRIBUTION AND ASK FOR THE CODE NAME DESIRED FOR THE OVERALL DISTRIBUTION. (AT THIS POINT, THE COMPUTER WILL ALSO WRITE OUT THE CODE NAMES OF ALL OF THE FILES INCLUDED IN THAT TOTAL.)

```
*WHAT 6-SPACE CODE SHOULD WE USE FOR
*THE OVERALL GROUP TOTAL?
*   SUM OF: (FILE CODE NAMES, UP TO 8 FILES)
```

IF THE "OVERALL TOTAL" OPTION HAS BEEN SELECTED, COMPUTER ASKS:

```
*WRITE OUT (NAME) DATA:
*     (1) AS TYPED TABLE?
*     (2) INTO COMPUTER FILE?
*     (3) BOTH? OR
*     (4) NO WRITE-OUT WANTED. (ANS 1-4)
```

IF "1", LISTS CODE NAME OF DISTRIBUTION, TYPES OUT THE CODE NAMES OF THE INDIVIDUAL GROUPS INCLUDED IN THE TOTAL DISTRIBUTION, AND TYPES OUT THE COMPLETE 8 X 10 MATRIX, INCLUDING COLUMN- AND ROW-VALUES.

IF "2", THE FILE CONTAINING THE TOTAL DISTRIBUTION IS WRITTEN INTO A COMPUTER STORAGE AREA. AT THE END OF THE PROGRAM RUN THE COMPUTER WILL LIST THOSE FILES WHICH HAVE BEEN WRITTEN INTO COMPUTER STORAGE DURING THE CURRENT RUN. THESE FILES MAY THEN BE SAVED OR DELETED AS THE OPERATOR SO DESIRES.

IF "3", THE COMPUTER WILL BOTH TYPE OUT THE COMPLETE FILE AND ENTER IT INTO STORAGE. UNDER "2" AND "3", THE COMPUTER TYPES :

```
*COMPUTER FILE ENTERED
```

IF OPTION "4" IS SELECTED, THE COMPUTER SKIPS THE ABOVE AND GOES TO THE NEXT STEP IN THE "OVERALL TOTAL" OPTION AND WRITES :

```
*COMPUTED DATA FOR TOTAL DISTRIBUTION (NAME)
*   SUM OF: (NAMES OF FILES MAKING UP TOTAL)
*
*     EQUATION:   LOG Y = (A) + (B)LOG X    R= (COEF. OF CORR.)
*
*     BASE N = (NO.)     BASE VARIANCE = (VALUE)
*

*IS A 35-YEAR CAREER PROJECTION DESIRED? (Y OR N)
```

IF "Y" IS GIVEN, PROGRAM ITERATES THE ABOVE GRADE/TIME EQUATION 35 TIMES, BEGINNING WITH THE BASE POINT USED IN FIRST COLUMN. IF "N", PROGRAM PROCEEDS TO NEXT STEP.

UNDER THE "STD PATTERN" OPTION MENTIONED ON PAGE 3, SOME OF THE VARIABLES DISCUSSED ABOVE CAN BE SET AT THE START OF THE RUN AND THE COMPUTER WILL NOT PRINT THE QUESTIONS PERTAINING TO THOSE VARIABLES. THE VARIABLES INVOLVED ARE :

(1) COL-VALUES -- WILL BE SET TO OPTION "1"
(2) ROW-VALUES -- OPERATOR IS STILL GIVEN A CHOICE BETWEEN GS-5/15 AND GS-1/8
(3) SUBGROUP TESTS -- WILL BE SET TO OPTION "1"
(4) COMPUTED DATA -- WILL NOT BE PRINTED OUT
(5) 35-YEAR PROJECTION -- WILL BE SET TO "N"

QUESTIONS PERTAINING TO OTHER VARIABLES WILL STILL BE WRITTEN OUT.

* = COMPUTER WRITTEN

NEXT: IF EITHER OPTION "1" OR "2" UNDER "SUBGROUP TESTS" HAS BEEN CHOSEN, THE COMPUTER ASKS :

```
*WRITE OUT TEST QUESTIONS? (Y OR N)
```

IF "Y", THE COMPUTER WRITES OUT THE QUESTIONS ANSWERED BY THE STATISTICAL TESTS CONTAINED IN THE PROGRAM :

```
*SUBGROUPS WERE TESTED FOR THE FOLLOWING QUESTION(S):
*
*     QUESTION (1):
*
*        IS THE GRADE-ADVANCEMENT CURVE OF EACH WORKFORCE SUBGROUP
*        SIGNIFICANTLY DIFFERENT (HIGHER, LOWER) THAN THE NORM FOR
*        THEIR OVERALL OCCUPATIONAL GROUP (NAME)?
*
*     QUESTION (2):
*
*        IS THE SPREAD OF SUBGROUP EMPLOYEE GRADE (SKILL) LEVELS
*        ABOVE AND BELOW THIS NORMAL GRADE-ADVANCEMENT CURVE
*        SIGNIFICANTLY DIFFERENT FROM THE (NAME) NORM?
*
*     QUESTION (3):
*
*        DOES THE OVERALL GRADE/TIME PATTERN RESULTING FROM:
*
*           (A) THE SUBGROUP GRADE-ADVANCEMENT CURVE; AND
*           (B) THE SUBGROUP GRADE (SKILL) SPREAD
*
*        SIGNIFICANTLY DIFFER FROM THAT OF THE (NAME) NORM?
```

THE COMPUTER THEN PRINTS OUT THE FOLLOWING PARAGRAPH EXPLAINING THE STATISTICAL SIGNIFICANCE OF TEST RESULTS :

```
*TEST DIFFERENCES ARE CONSIDERED STATISTICALLY SIGNIFICANT WHEN THE
*PROBABILITY THAT THEIR OCCURRENCE COULD BE ATTRIBUTED TO CHANCE
*IS 0.05 (1 CHANCE IN 20) OR LESS.
```

IF "N", THE COMPUTER GOES DIRECTLY TO THE WRITE OUT OF THE TABLE WHICH SUMMARIZES THE RESULTS OF THE STATISTICAL TESTS (I.E., ANSWERS THE ABOVE QUESTIONS) MADE IN THE PROGRAM. (THIS TABLE IS PRINTED OUT IN ALL CASES WHERE SUBGROUP DIFFERENCES ARE TESTED.)

THE THREE MAIN COLUMN HEADINGS OF THIS TABLE REFER TO THE THREE QUESTIONS TESTED AS FOLLOWS :

QUESTION #	COLUMN HEADING
(1)	ADVANCEMENT CURVES
(2)	GRD (SKILL) DISTRIB
(3)	OVERALL G/T PATTERN

AFTER PRINTING OUT THE "SUMMARY OF TEST RESULTS", THE COMPUTER ASKS:

```
*WRITE OUT DETAILED TEST DATA? (Y OR N)
```

IF "Y", THE MACHINE PRINTS A TABLE CONTAINING THE STATISTICAL DATA WHICH WAS USED IN THE STATISTICAL TESTS IN THE PROGRAM. DATA ARE PRINTED OUT FOR BOTH THE OVERALL DISTRIBUTION (OR OUTSIDE GROUP) AND EACH OF THE PREVIOUSLY-ENTERED SUBGROUPS.

IF "N", (OR, IF "Y", AFTER THE PRINTING OUT OF THE "SUMMARY OF THE STATISTICAL DATA"), THE COMPUTER WRITES :

```
*NEXT: FILE WRITE-OUT:
*
*ENTER NO. OF FILE(1-8) OR
*9 (NO FILE WANTED)
```

THIS OPTION ALLOWS THE OPERATOR TO WRITE OUT ANY FILE HE MAY WISH TO SEE FOR MORE DETAILED STUDY. "FILE(1-8)" REFERS TO THE FILES ENTERED AT THE BEGINNING OF THE PROGRAM IN THE ORDER THEY WERE ENTERED. (FILES ARE ALSO LISTED BY NUMBER IN THE "SUMMARY OF STATISICAL DATA" TABLE.) AN ENTRY OF "1" CAUSES THE FIRST FILE TO BE PRINTED, "2" THE SECOND, AND SO ON. THESE NUMBERS MAY BE ENTERED IN ANY ORDER.

IF A "9" IS ENTERED, NO FURTHER FILES ARE WRITTEN OUT AND THE COMPUTER THEN WRITES:

```
*     RUN COMPLETED. NEXT:
*(1) AGAIN (AUTO)? (2) AGAIN (STEP BY STEP)? OR (3) QUIT. (1,2 OR 3)
```

IF "1" IS GIVEN, THE PROGRAM RUNS NEW FILES THROUGH EXACTLY THE SAME PROGRAM-OPTION STEPS WHICH THE OPERATOR CHOSE IN THE PREVIOUS RUN-THROUGH WITHOUT ANY OPERATOR PARTICIPATION OTHER THAN ENTERING THE NEW FILE CODES. IF "2" IS GIVEN, THE PROGRAM AGAIN STEPS THROUGH EACH OF THE ABOVE STEPS (EXCEPT THAT OPTION IS GIVEN TO RETAIN THE ABOVE-SELECTED COLUMN- AND GRADE- PATTERNS). IF "3" IS GIVEN, THE COMPUTER LISTS THE NAMES OF FILES WHICH HAVE BEEN PLACED IN COMPUTER STORAGE AS FOLLOWS :

```
*(NO.) FILES ENTERED DURING RUN;
*"SAVE" THOSE TO BE KEPT:
*   (NAME1)
*   (NAME2)
      ETC.
```

THE OPERATOR MAY THEN SAVE ANY FILES HE WISHES TO USE AGAIN. AFTER THIS LISTING, THE RUN IS TERMINATED.

NOTE: FILES STORED DURING A PROGRAM RUN CANNOT BE ACCESSED DURING THAT RUN. TO USE THESE FILES, OPERATOR MUST TERMINATE THE CURRENT PROGRAM RUN AND BEGIN A NEW RUN.

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

GS810

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

THIS PROGRAM ANALYZES 10-YEAR GRADE TRENDS IN 1-8 GROUPS OF THREE OR MORE EMPLOYEES. COMPUTES 10-YEAR GRADE/TIME CURVE AND VARIANCE (OR ACCEPTS THE GRADE/TIME EQUATION AND VARIANCE OF AN OUTSIDE GROUP) AND TESTS THE GRADE TREND OF EACH ENTERED GROUP FOR STATISTICALLY SIGNIFICANT DIFFERENCES FROM THE TREND OF THE OVERALL (OR OUTSIDE) GROUP. ENTERED GROUPS MUST EACH HAVE A POPULATION OF THREE OR MORE FOR VALID RESULTS TO BE OBTAINED. (FOR AN OUTSIDE GROUP'S EQUATION AND VARIANCE TO BE USED AS A BASIS FOR ANALYZING GROUPS ENTERED IN THIS PROGRAM, THE DATA FOR THE OUTSIDE GROUP MUST HAVE INCLUDED ALL -- REPEAT, ALL -- OF THE INDIVIDUAL GROUP DATA ENTERED HERE. THESE GROUPS, THAT IS, MUST BE COMPONENTS OF THE OUTSIDE GROUP.)

DATA ARE ENTERED INTO THE PROGRAM FROM PREVIOUSLY-STORED FILES ACCESSED VIA A 6-SPACE CODE (EXPLAINED IN THE OPERATION MANUAL). FILES CONSIST OF 8 CONSECUTIVE TYPED LINES OF 10 ENTRIES EACH, SEPARATED BY COMMAS, WITH EACH ENTRY MADE UP OF 1-4 INTEGERS (OR ZEROES -- NO BLANKS).

REQUIRED SUBPROGRAMS: RLL810, VLL810, FTEST, ANDXP

```
      DIMENSION IARRAY(8,10,8),WORK(10,10),NARRAY(8),
      1  X(10),Y(8),ITR(8,10),SIG(3,8),P(3),
      1  NXX(36),SUMN(8),VARGP(8),NOTEST(8),PATDIF(8),
      1  N(8),R(8),A(8),B(8),VAROWN(8),DY(8),
      1  DIFF(8),RANGE(8),NDASH(36),NAMFIL(20)
      LOGICAL IMP,INP,IP,IQ,NOTEST,ISTD
      DOUBLE PRECISION DIFF,PATDIF,SIG,NARRAY,KARRAY,IDBASE
      1   ,RANGE,NAMFIL
      DATA INP/.TRUE./,IQ/.FALSE./,IV/'N'/,IMP/.FALSE./,ITIMES/20/,
      1  (NXX(I),I=1,36)/36*'XX'/,(NDASH(I),I=1,36)/36*'--'/,NSAVED/0/
      1  ,MT/'Y'/,ISTD/.FALSE./,(DY(I),I=1,8)/5.25,5.75,6.5,7.5,
      1  8.25,8.75,9.25,9.75/
120   WRITE (5,130)
130   FORMAT (//1X,'THIS PROGRAM ANALYZES GRADE/TIME TRENDS IN 1-8')
      WRITE (5,150)
150   FORMAT (1X,'GROUPS OF 3 OR MORE EACH; USES 8-GRADE X 10-YEAR'
      1  /1X,'DATA ARRAYS.')
      WRITE (5,210)
210   FORMAT (/1X,'ENTER FILES VIA 6-SPACE CODE (SEE INSTR).')
220   WRITE (5,230)
230   FORMAT (///1X,'HOW MANY FILES DO YOU WISH TO ANALYZE?'
      1  ' (ANS 1-8)')
      READ (5,260)NF
260   FORMAT (I2)
      WRITE (5,280)
280   FORMAT (1X,'PLEASE ENTER CODE NAME OF EACH FILE:')
      DO 340 I=1,NF
      WRITE (5,310)I
310   FORMAT (1X,I2,'=')
      READ (5,330)NARRAY(I)
330   FORMAT (A6)
340   CONTINUE
      DO 410 I=1,NF
370   KARRAY=NARRAY(I)
      CALL IFILE(20,KARRAY)
      READ (20,*)((IARRAY(I,J,K),J=1,10),K=1,8)
400   FORMAT (10I4,7(/,10I4))
410   CONTINUE
      WRITE (5,430)
430   FORMAT (//1X,'DATA TAKEN FROM FILES.  NEXT:')
      IF(INP)WRITE(5,450)
 50   FORMAT(/1X,'"STD PATTERN" DESIRED? (Y OR N)')
      IF(INP)READ(5,560)IE
      IF(IE.EQ.'N')ISTD=.FALSE.;IF(IE.EQ.'N')GO TO 530
      IF(INP)WRITE(5,490)
490   FORMAT(/1X,'(1) 5/15 OR (2) 1/8? (1 OR 2)')
      IF(INP)READ(5,260)IR
      IF(IR.EQ.1)NYPAT=1;IF(IR.EQ.2)NYPAT=2
      NXPAT=1;LOP=1;MOR='N';ISTD=.TRUE.
530   IF ((INP).AND.(IMP).AND..NOT.ISTD) WRITE (5,540)
540   FORMAT (/1X,'SAME GRADE(WT) AND COL-VALUES PATTERN? (Y OR N)')
      IF((INP).AND.(IMP).AND..NOT.ISTD) READ (5,560)IV
```

```
560     FORMAT (A2)
        IF((INP).AND.(IMP).AND.IV.EQ.´Y´) GO TO 1250
        IF(((INP).OR.(ISTD)).AND.IV.EQ.´N´) GO TO 600
        GO TO 1330
600     IF(.NOT.ISTD)WRITE (5,610)
610     FORMAT (/6X,´(A) SELECT COLUMN-VALUE PATTERN WANTED:´)
        IF(.NOT.ISTD)WRITE (5,630)
630     FORMAT (10X,´(USE COL. MIDPOINTS: "0-1" YRS="0.5", ETC.)´)
        IF(.NOT.ISTD)WRITE (5,650)
650     FORMAT (/10X,´(1) 0.5,1.5, ETC.´
        1  /10X,´(2) 1.,2., ETC.´
        1  /10X,´(3) OTHER (1,2 OR 3)´)
        IF(.NOT.ISTD)READ (5,260) NXPAT
        IF(NXPAT.EQ.3) GO TO 760
        IF (NXPAT.EQ.1)EI=-0.5
        IF(NXPAT.EQ.2)EI=0.0
        DO 740 I=1,10
        EX=EI+I
740     X(I)=EX
        GO TO 830
760     WRITE (5,770)
770     FORMAT (6X,´ENTER DESIRED VALUE FOR MIDPOINT OF COL.:´)
        DO 820 I=1,10
        WRITE(5,1040)I
        READ(5,810)X(I)
810     FORMAT(F6.3)
820     CONTINUE
830     IF(.NOT.ISTD)WRITE (5,840)
840     FORMAT(/6X,´(B) SELECT GRADE PROGRESSION´
        1  /10X,´(ROW-VALUE) PATTERN WANTED:´)
        IF(.NOT.ISTD)WRITE(5,870)
870     FORMAT(/10X,´(1) GS-5/15´
        1  /10X,´(2) GS-1/8´
        1  /10X,´(3) OTHER (1,2 OR 3)´)
        IF(.NOT.ISTD)READ(5,260)NYPAT
        IF(NYPAT.EQ.3) GO TO 990
        IF(NYPAT.EQ.1)EI=7.5
        IF(NYPAT.EQ.2)GO TO 962
        DO 960 I=1,8
        EY=EI+I
960     Y(I)=EY
        IF(ISTD)GO TO 1250
        GO TO 1080
962     DO 963 I=1,8
963     Y(I)=DY(I)
        IF(ISTD) GO TO 1250
        GO TO 1080
990     WRITE(5,1000)
1000    FORMAT(/6X,´ENTER MIDPOINT (E.G.,GRD."1"=(1.-1.99)="1.5")´
        1  ´ FOR ROW:´)
        DO 1060 I=1,8
        WRITE(5,1040)I
1040    FORMAT(6X,I2)
        READ(5,810)Y(I)
1060    CONTINUE
```

```
      GO TO 1250
1080  IF (NYPAT.NE.1) GO TO 1250
      WRITE (5,1090)
1090  FORMAT (/1X,´PATTERN APPLIED:´)
      WRITE (5,1120)
1120  FORMAT (/1X,´(SO THAT "2-GRADE-INTERVAL-SERIES" GRADES ´
      1  ´PROGRESS AT DHE SAME´)
      WRITE (5,1150)
1150  FORMAT (1X,´"1-GRADE/1-STEP" RATE BELOW GS-11´
      1  ´ AS THEY DO ABOVE GS-11, GS-5/9´)
      WRITE (5,1180)
1180  FORMAT (1X,´GRADES WERE WEIGHTED: GS-5="8", GS-7="9", AND´
      1  ´ GS-9="10".)´)
1200  WRITE (5,1210)
1210  FORMAT (/1X,´(0.50 WAS ADDED TO EACH GRADE-VALUE AS THE MID´
      1  ´POINT OF THE´)
      WRITE (5,1240)
1240  FORMAT (1X,´TOTAL RANGE OF JOBS IN THAT GRADE.)´)
1250  IF((INP).AND..NOT.ISTD)WRITE (5,1260)
1260  FORMAT(//1X,´TEST SUBGROUPS AGAINST:´,/6X,´(1) THEIR OWN ´
      1  ´SUM?´,/6X,´(2) AN OUTSIDE GROUP?´)
      IF((INP).AND..NOT.ISTD)WRITE(5,1290)
1290  FORMAT(6X,´(3) NO TESTS DESIRED.  (1,2 OR 3)´)
      IF((INP).AND..NOT.ISTD) READ(5,260)LOP
      IF(LOP.EQ.3)MT=´N´
      IF(MT.EQ.´N´) GO TO 1520
1330  IF (LOP.EQ.1.OR.MT.EQ.´N´) GO TO 1520
1340  WRITE(5,1350)
1350  FORMAT (/1X,´WHAT 6-SPACE CODE NAME SHOULD WE USE FOR THE´
      1  /1X,´OUTSIDE (BASE) GROUP?´)
      READ(5,1380)IDBASE
1380  FORMAT(A6)
      WRITE (5,1400)IDBASE
1400  FORMAT(1X,´PLEASE ENTER THE FOLLOWING FOR BASE GROUP´
      1  ´ ´,A6,´:´)
      WRITE (5,1430)IDBASE
1430  FORMAT(/6X,A6,´ GRADE/TIME CURVE EQUATION´
      1  /5X,´(E.G., 1.12345,.123456): ´)
      READ (5,*)YIN,SL
1540  FORMAT (2F9.6)
      WRITE (5,1480)IDBASE
1480  FORMAT (6X,A6,´ N, VARIANCE (E.G.: 123,1.12345):´)
      READ (5,*) NBAS,VARBAS
1500  FORMAT (I5,F9.6)
      GO TO 2280
1520  WRITE(5,1530)
1530  FORMAT (/1X,´WHAT 6-SPACE CODE SHOULD WE USE FOR´
      1  /1X,´THE OVERALL GROUP TOTAL?´)
      WRITE (5,1560)(NARRAY(I),I=1,NF)
1560  FORMAT (4X,´SUM OF:´,8(1X,´ ´,A6))
      READ (5,330)IDBASE
      IF (INP) WRITE(5,1590)IDBASE
1590  FORMAT (/1X,´WRITE OUT ´,A6,´ DATA:´
      1  /6X,´(1) AS TYPED TABLE? ´,/6X,´(2) INTO COMPUTER FILE?´)
      IF(INP)WRITE(5,1620)
```

```
1620      FORMAT(6X,'(3) BOTH? OR'
          1  /6X,'(4) NO WRITE-OUT WANTED. (ANS 1-4)')
          IF(INP) READ (5,260)LP
          DO 1670 I=1,8
          DO 1670 J=1,10
1670      WORK(I,J)=0.0
          DO 1740 I=1,NF
          DO 1740 JJ=1,10
          DO 1740 II=1,8
          WA=IARRAY(I,JJ,II)
          WORK(II,JJ)=WORK(II,JJ)+WA
          ITR(II,JJ)=WORK(II,JJ)
1740      CONTINUE
          CALL RLL810(X,Y,WORK,YIN,SL,NBAS,IER)
          NBASDF=NBAS-1
          CALL VLL810(YIN,SL,X,Y,WORK,VARBAS,VARTOT,RBAS)
          IF(LP.EQ.4)GO TO 2020
          IF(LP.EQ.2)GO TO 1920
          WRITE(5,1810)IDBASE
1810      FORMAT(///1X,'CUMULATED DISTRIBUTION ',A6,':')
          WRITE(5,1560)(NARRAY(K),K=1,NF)
          WRITE (5,1840) (X(I),I=1,10)
1840      FORMAT (/1X,8X,10F6.2)
          WRITE(5,1860)
1860      FORMAT(2X,'GRD(WT)')
          II=9
          DO 1900 IJ=1,8
          I=II-IJ
          WRITE(5,1890)Y(I),(ITR(I,J),J=1,10)
1890      FORMAT(2X,F5.2,2X,10I6)
1900      CONTINUE
          IF(LP.EQ.1) GO TO 2020
1920      ITIMES=ITIMES+1
          IF(ITIMES.GT.24)ITIMES=21
          CALL OFILE(ITIMES,IDBASE)
          WRITE(ITIMES,400)((ITR(I,J),J=1,10),I=1,8)
C         IF CALL SAVE OPTION IS DESIRED, REMOVE * FROM NEXT 3 LINES
C         CALL SAVE(ITIMES,ISTAT)
C         IF (ISTAT.EQ.0) WRITE (5,1933)
C         1933 FORMAT (//1X,'FILE SAVED')
          WRITE(5,1970)
1970      FORMAT(//1X,'COMPUTER FILE ENTERED.')
2020      IF(ISTD)GO TO 2280
1990      WRITE (5,2000)IDBASE
2000      FORMAT (///1X,'COMPUTED DATA FOR TOTAL DISTRIBUTION ',
          1  A6,':')
          WRITE (5,1560)(NARRAY(I),I=1,NF)
          WRITE (5,2080)YIN,SL,RBAS
2080      FORMAT(/6X,'EQUATION:  LOG Y  =',F9.6,' +',F9.6,' LOG X'
          1  '    R=',F9.6)
          WRITE (5,2110)NBAS,VARBAS
2110      FORMAT (/6X,'BASE N =',I4,'    BASE VARIANCE = ',F9.6)
2120      IF((INP).AND..NOT.ISTD)WRITE(5,2130)
2130      FORMAT(//1X,'35-YEAR (CAREER) PROJECTION? (Y OR N)')
          IF((INP).AND..NOT.ISTD)READ(5,560)MCR
```

```
      IF (MOR.EQ.'N') GO TO 2280
      WRITE (5,2170)IDBASE
2170  FORMAT (//1X,'35-YEAR CAREER PROJECTION OF ',A6)
      WRITE (5,1560)(NARRAY(I),I=1,NF)
      WRITE (5,2200)
2200  FORMAT (/1X,' YEAR   GRADE',/)
      DO 2270 I=1,35
      XI=I-1
      XI=XI+X(1)
      PY=10.**(YIN+SL*ALOG10(XI))
      WRITE (5,2260)XI,PY
2260  FORMAT (1X,F5.2,F9.4)
2270  CONTINUE
2280  IF (MT.EQ.'N') GO TO 3970
      DO 2960 K=1,NF
      DO 2310 L=1,3
2310  P(L)=.99
      NOTEST(K)=.FALSE.
      DIFF(K)='    '
      PATDIF(K)='     '
      RANGE(K)='      '
      SUMN(K)=0.0
      DO 2410 J=1,10
      DO 2410 I=1,8
      WORK(I,J)=IARRAY(K,J,I)
      SUMN(K)=SUMN(K)+WORK(I,J)
2410  CONTINUE
      IF(SUMN(K).LT.3.) NOTEST(K)=.TRUE.
      IF(NOTEST(K)) GO TO 2880
      CALL VLL810(YIN,SL,X,Y,WORK,VAR2,VTOT,RR)
      F=VAR2/VARBAS
      VARGP(K)=VAR2
      NGMS=SUMN(K)-1.
      CALL FTEST (NGMS,NBASDF,F,PPAT)
      P(1)=PPAT
      PATDIF(K)=' ---- '
      IF(PPAT.GT.0.05) GO TO 2530
      PATDIF(K)=' DIFF.'
2530  CALL RLL810(X,Y,WORK,AA,BB,NSG,IER)
      IF(IER.EQ.0) NOTEST(K)=.TRUE.
      IF(IER.EQ.0) PATDIF(K)='     '
      IF(IER.EQ.0) P(1)=.99
      IF (IER.EQ.0) GO TO 2880
      CALL VLL810(AA,BB,X,Y,WORK,VSG,VARS,RS)
      N(K)=NSG
      R(K)=RS
      A(K)=AA
      B(K)=BB
      VAROWN(K)=VSG
      IF(VAROWN(K).EQ.0.0) NOTEST(K)=.TRUE.
      IF (NOTEST(K)) GO TO 2880
      FF=VAR2/VSG
      NSGDF=NSG-1
      CALL FTEST (NSGDF,NSGDF,FF,PP)
      DIFF(K)=' ---- '
```

```
        P(2)=PP
        IF(PP.GT.0.05) GO TO 2750
        DIFF(K)=' DIFF.'
        IF(YIN.GT.AA.AND.SL.GT.BB) DIFF(K)= 'LOWER '
        IF(AA.GT.YIN.AND.BB.GT.SL) DIFF(K)='HIGHER'
2750    FD=VARBAS/VSG
        RANGE(K)='  ----  '
        IF (FD.LE.1.) GO TO 2830
        CALL FTEST(NBASDF,NSGDF,FD,PD)
        P(3)=PD
        IF(PD.GT.0.05) GO TO 2880
        RANGE(K)=' SMALLER'
        GO TO 2880
2830    FD=1./FD
        CALL FTEST(NSGDF,NBASDF,FD,PD)
        P(3)=PD
        IF(PD.GT.0.05) GO TO 2880
        RANGE(K)=' GREATER'
2880    DO 2960 I=1,3
        SIG(I,K)='      '
        IF(.NOT.NOTEST(K))SIG(I,K)=' ---- '
        IF(P(I).GT.0.05) GO TO 2960
        IF(P(I).LE.0.05)SIG(I,K)='<=.05 '
        IF(P(I).LE.0.01)SIG(I,K)='<=.01 '
        IF(P(I).LE..001)SIG(I,K)='<=.001'
        IF(P(I).LE..0001)SIG(I,K)='<.0001'
2960    CONTINUE
        WRITE(5,2980)(NXX(I),I=1,35)
2980    FORMAT(//1X,35A2)
        IF(INP) WRITE(5,3000)
3000    FORMAT (/1X,'WRITE OUT TEST QUESTIONS? (Y OR N)')
        IF(INP) READ (5,560) IWQ
        IF(IWQ.EQ.'N') GO TO 3450
3030    WRITE(5,3040)
3040    FORMAT(//1X,'SUBGROUPS WERE TESTED FOR THE FOLLOW'
        1  'ING QUESTION(S):',//6X,'QUESTION (1):')
        WRITE(5,3070)
3070    FORMAT(/9X,'IS THE GRADE-ADVANCEMENT CURVE OF EACH '
        1  'WORKFORCE SUBGROUP')
        WRITE(5,3100)
3100    FORMAT(9X,'SIGNIFICANTLY DIFFERENT (HIGHER, LOWER) THAN'
        1  ' THE NORM FOR')
        WRITE(5,3130)IDBASE
3130    FORMAT(9X,'THEIR OVERALL OCCUPATIONAL GROUP (',A6,')?')
3140    WRITE(5,3150)
3150    FORMAT(/6X,'QUESTION (2):')
        WRITE(5,3170)
3170    FORMAT(/9X,'IS THE SPREAD OF SUBGROUP EMPLOYEE'
        1  ' GRADE (SKILL) LEVELS')
        WRITE(5,3200)
3200    FORMAT(9X,'ABOVE AND BELOW THIS NORMAL GRADE-ADVANCEMENT'
        1  ' CURVE')
        WRITE (5,3230)IDBASE
3230    FORMAT(9X,'SIGNIFICANTLY DIFFERENT FROM THE ',A6,' NORM?')
        WRITE(5,3250)
```

```
3250    FORMAT(/6X,'QUESTION (3):')
3260    WRITE(5,3270)
3270    FORMAT(/9X,'DOES THE OVERALL GRADE/TIME PATTERN '
        1  'RESULTING FROM:')
        WRITE(5,3300)
3300    FORMAT(/12X,'(A) THE SUBGROUP GRADE-ADVANCEMENT CURVE;'
        1  ' AND')
        WRITE(5,3330)
3330    FORMAT(12X,'(B) THE SUBGROUP GRADE (SKILL) SPREAD')
        WRITE(5,3350)IDBASE
3350    FORMAT(/9X,'SIGNIFICANTLY DIFFER FROM THAT OF THE '
        1  ,A6,' NORM?')
        WRITE(5,3380)
3380    FORMAT(/1X,'TEST DIFFERENCES ARE CONSIDERED STATISTICALLY'
        1  ' SIGNIFICANT WHEN THE')
3400    WRITE(5,3410)
3410    FORMAT(1X,'PROBABILITY THAT THEIR OCCURRENCE COULD BE '
        1  'ATTRIBUTED TO CHANCE')
        WRITE(5,3440)
3440    FORMAT(1X,'IS 0.05 (1 CHANCE IN 20) OR LESS.')
3450    WRITE(5,3460)
3460    FORMAT(//1X,'KEY:',/4X,'"    " = NO TEST (N<3)',
        1  /4X,'"----" = NOT SIG.')
        WRITE(5,3490)(NDASH(I),I=1,11),(NDASH(I),I=1,11)
3490    FORMAT(//1X,11A2,1X,'SUMMARY OF TEST RESULTS',1X,'-',11A2)
        WRITE(5,3510)
3510    FORMAT(/8X,'ADVANCEMENT CURVES   GRD (SKILL) DISTRIB '
        1  '   OVERALL G/T PATTERN')
        WRITE(5,3550)(NDASH(I),I=1,9),(NDASH(I),I=1,10),(NDASH(I),
        1  I=1,10)
3550    FORMAT(8X,9A2,'-',2X,10A2,2X,10A2)
        WRITE(5,3570)
3570    FORMAT(1X,'FILE   SUBGROUP  P. DIFF.S  SUBGROUP   P. DIFF.S'
        1  '   SUBGROUP   P. DIFF.S')
        WRITE(5,3600)
3600    FORMAT(1X,'CODE   CURVE IS  ATTR CHCE  SPRD IS    ATTR CHCE'
        1  '   PTTRN IS   ATTR CHCE')
        WRITE(5,3630)(NDASH(I),I=1,9),(NDASH(I),I=1,10),(NDASH(I),I=1,10)
3630    FORMAT(1X,'------ ',9A2,'-',2X,10A2,2X,10A2,/)
        DO 3680 K=1,NF
        WRITE (5,3670) NARRAY(K),DIFF(K),SIG(2,K),
        1  RANGE(K),SIG(3,K),PATDIF(K),SIG(1,K)
3670    FORMAT(1X,A6,2X,A6,5X,A6,3X,A8,5X,A6,4X,A6,6X,A6)
3680    CONTINUE
        WRITE (5,3700)(NDASH(I),I=1,35)
3700    FORMAT(/1X,35A2)
        IF(INP) WRITE(5,3720)
3720    FORMAT (//1X,'WRITE OUT DETAILED TEST DATA? (Y OR N)')
        IF(INP) READ (5,560)IWD
        IF(IWD.EQ.'N')GO TO 3970
        WRITE(5,3760)(NDASH(I),I=1,8),IDBASE,(NDASH(I),I=1,9)
3760    FORMAT(///1X,8A2,' SUMMARY OF ',A6,' STATISTICAL DATA ',9A2)
        WRITE(5,3780)
3780    FORMAT(/,27X,'GRADE/TIME EQUATION',10X,'VARIANCE FROM:')
        WRITE(5,3800)(NDASH(I),I=1,12),(NDASH(I),I=1,8)
```

```
3800     FORMAT(1X,'GP.  FILE',5X,'GP. ',5X,12(A2),'-',5X,'-',8(A2))
         WRITE(5,3820)
3820     FORMAT(1X,'NO.  CODE      POP.       (A)       (B)       (R)'
         1 '             GP. EQ.  OWN EQ.')
         WRITE(5,3850)(NDASH(I),I=1,9),(NDASH(I),I=1,12),(NDASH(I),I=1,8)
3850     FORMAT(1X,9A2,5X,12A2,'-',5X,'-',8A2,/)
         WRITE(5,3870)IDBASE,NBAS,YIN,SL,RBAS,VARBAS
3870     FORMAT(3X,A6,4X,I6,F12.5,F9.5,F9.5,6X,'(NA)',F12.5,/)
         DO 3950 K=1,NF
         IF(.NOT.NOTEST(K))WRITE(5,3910)K,NARRAY(K),N(K),A(K),
         1  B(K),R(K),VARGP(K),VAROWN(K)
3910     FORMAT (1X,I2,':',2X,A6,1X,I6,F12.5,2F9.5,F12.5,F10.5)
         IF(NOTEST(K)) N(K)=SUMN(K)
         IF(NOTEST(K))WRITE(5,3940)K,NARRAY(K),N(K)
3940     FORMAT(1X,I2,':',2X,A6,1X,I6)
3950     CONTINUE
         WRITE(5,3700)(NDASH(I),I=1,35)
3970     IF(LP.EQ.1.OR.LP.EQ.4.OR.LOP.EQ.2) GO TO 4010
         NSAVED=NSAVED+1
4000     NAMFIL(NSAVED)=IDBASE
4010     IF(MT.EQ.'N') GO TO 4150
4020     WRITE(5,4030)
4030     FORMAT(//1X,'NEXT: FILE WRITE-OUT:',//1X,'ENTER NO. OF FILE'
         1  '(1-8) OR',/1X,'9 (NO FILE WANTED)')
         READ(5,260)NFP
         IF (NFP.EQ.9) GO TO 4150
         WRITE(5,4080)NARRAY(NFP)
4080     FORMAT(///1X,A6,' FILE:')
         WRITE(5,1840)(X(I),I=1,10)
         WRITE(5,1860)
         II=9
         DO 4130 IJ=1,8
         I=II-IJ
         WRITE(5,1890) Y(I),(IARRAY(NFP,J,I),J=1,10)
4130     CONTINUE
         GO TO 4020
4150     WRITE(5,4160)(NXX(I),I=1,34)
4160     FORMAT(//1X,34A2)
4170     WRITE(5,4180)
4180     FORMAT(///,6X,'RUN COMPLETED. NEXT:')
4190     WRITE(5,4200)
4200     FORMAT(1X,'(1) AGAIN (AUTO)? (2) AGAIN (STEP BY STEP)? OR'
         1  ' (3) QUIT. (1,2 OR 3)')
         READ(5,260)LGN
         IF(LGN-2) 4240,4270,4320
4240     INP=.FALSE.
         IQ=.TRUE.
         GO TO 220
4270     INP=.TRUE.
         IQ=.FALSE.
         IMP=.TRUE.
         MT='Y'
         GO TO 220
4320     IF(NSAVED.NE.0)WRITE(5,4340)NSAVED,(NAMFIL(NU),
         1  NU=1,NSAVED)
```

```
4340    FORMAT(//1X,I2,´ FILES ENTERED DURING RUN;´
        1  /,´*SAVE* THOSE TO BE KEPT:´,/,(3X,A6))
        STOP
        END
```

```
      SUBROUTINE RLL810(X,Y,TABLE,A,B,N,IER)
C     THIS PROGRAM COMPUTES LOG(Y)=A+B(LOG)X EQUATION FOR DATA
C     ENTERED IN 8(Y) BY 10(X) MATRIX.  RETURNS A AND B IN LOG FORM.
      DIMENSION X(10),Y(8),TABLE(10,10)
      IER=1
      SUMX=0.0
      SUMY=0.0
      SUMXY=0.0
      SUMX2=0.0
      SUMN=0.0
      NX=0
      NCOL=0
      DO 18 J=1,10
      DO 15 I=1,8
      IF(TABLE(I,J).NE.0.0)NCOL=NCOL+1
      SUMX=SUMX+ALOG10(X(J))*TABLE(I,J)
      SUMY=SUMY+ALOG10(Y(I))*TABLE(I,J)
      SUMXY=SUMXY+(ALOG10(X(J))*ALOG10(Y(I)))*TABLE(I,J)
      SUMX2=SUMX2+(ALOG10(X(J))**2)*TABLE(I,J)
      SUMN=SUMN+TABLE(I,J)
15    CONTINUE
      IF(NCOL.NE.0)NX=NX+1
      NCOL=0
18    CONTINUE
      IF(NX.LT.2)IER=0
      IF(IER.EQ.0)GO TO 20
      EN=SUMN
      S1=EN*SUMX2-SUMX*SUMX
      S2=EN*SUMXY-SUMX*SUMY
      B=S2/S1
      A=SUMY/EN-(B*SUMX)/EN
      N=IFIX(EN)
      GO TO 30
20    A=0.0
      B=0.0
      N=IFIX(SUMN)
30    RETURN
      END
      SUBROUTINE VLL810 (A,B,X,Y,TABLE,VUNEXP,VARTOT,R)
C     ITERATES EQUATION OF FORM LOG(Y)=A+BLOG(X) FOR VALUES OF
C     X(1)-X(10) AND COMPUTES THE VARIANCE OF THE 8(Y) BY 10(X) TABLE
C     DATA FROM THIS LINE OF REGRESSION; COMPUTES THE TOTAL VARIANCE
C     OF THE TABLE DATA AND COMPUTES THE NON-LINEAR COEFFICIENT OF
C     CORRELATION (=SQRT(1-UNEXPLAINED VARIANCE/TOTAL VARIANCE)).
      DIMENSION X(10),Y(8),TABLE(10,10)
      SUMN=0.0
      SUNEXP=0.0
      SUMY=0.0
      SUMY2=0.0
      DO 16 J=1,10
      DO 16 I=1,8
      SUNEXP=SUNEXP+TABLE(I,J)*(Y(I)-10.**(A+B*ALOG10(X(J))))**2
      SUMY=SUMY+TABLE(I,J)*Y(I)
16    SUMN=SUMN+TABLE(I,J)
      YBAR=SUMY/SUMN
      DO 17 J=1,10
      DO 17 I=1,8
17    SUMY2=SUMY2+TABLE(I,J)*(Y(I)-YBAR)**2
      VARTOT=SUMY2/(SUMN-1.)
      VUNEXP=SUNEXP/(SUMN-1.)
      IF(VUNEXP.GE.VARTOT)R=0.0
      IF(VUNEXP.LT.VARTOT)R=SQRT(1.-VUNEXP/VARTOT)
      RETURN
      END
      SUBROUTINE FTEST(NGMS,NBAS,F,P)
C     APPROXIMATES THE PROBABILITY OF CHANCE OCCURRENCE OF THE
C     OBSERVED VALUE OF F, GIVEN NGMS (= N OF GROUP TESTED -- BY
C     DEFINITION, THE GREATER MEAN SQUARE) AND NBAS (= N OF THE
C     BASE GROUP).
C     REQUIRES SUBPROGRAM: FUNCTION ANDXP.
C     (REFERENCE: NBS, ABRAMOWITZ & STEGUN,AMS 55,9TH,26.6.15,P.947.)
      V1=NGMS-1
      V2=NBAS-1
      EX=1./3.
      F3=F**(EX)
      F23=F3**2
      B1=2./(9.*V1)
      B2=2./(9.*V2)
      T1=1.-B2
      T2=1.-B1
      TOP=F3*T1-T2
      BOT=SQRT(B1+F23*B2)
      X=TOP/BOT
      Q=ANDXP(X)
      P=1.-Q
      RETURN
      END
      FUNCTION ANDXP(X)
C     ACCUMULATIVE NORMAL DISTRIBUTION FUNCTION. APPROXIMATES P
C     FROM X (TO + OR - .0000001). (NBS-55, P. 932)
      X1 = ABS(X)
      T = 1./(1.+.2316419*X1)
      ANDXP = 1.-.3989423*EXP(-(X1**2)/2.)*(.3193815*T-.3565638*T**2
     1  +1.781478*T**3-1.821256*T**4+1.330274*T**5)
      IF (X.LT.0.0) ANDXP = 1.-ANDXP
      RETURN
      END
```

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

CS810

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

THE FOLLOWING PAGES SHOW THE COMPUTER ANALYSIS OF A GROUP OF CLERICAL OCCUPATIONS IN A SAMPLE AGENCY. THE OCCUPATIONS INVOLVED ARE:

GS-322	CLERK-TYPIST
GS-312	CLERK-STENOGRAPHER
GS-203	PERSONNEL CLERK
GS-318	SECRETARY

THE ANALYSIS PLAN USED IS AS FOLLOWS:

STEP	FILES USED	OUTPUT FILE
ANALYSIS OF BASIC FILES AND GENERATION OF SUMMARY FILES		
1	NF322,MF322	TOT322
2	NF312,MF312	TOT312
3	NF203,MF203	TOT203
4	NF318,MF318	TOT318
5	NF322,NF312,NF203,NF318	NFCL
6	MF322,MF312,MF203,MF318	MFCL
	(TERMINATE FIRST RUN)	
ANALYSIS OF SUMMARY FILES		
7	TOT322,TOT312,TOT203,TOT318	SUM
8	NFCL,MFCL	-----
	(TERMINATE SECOND RUN)	

```
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

WRITE OUT TEST QUESTIONS? (Y OR N)
N


KEY:
   "    " = NO TEST (N<3)
   "----" = NOT SIG.


---------------------- SUMMARY OF TEST RESULTS ----------------------

       ADVANCEMENT CURVES   GRD (SKILL) DISTRIB   OVERALL G/T PATTERN
       ------------------   -------------------   -------------------
FILE   SUBGROUP  P. DIFF.S  SUBGROUP   P. DIFF.S  SUBGROUP   P. DIFF.S
CODE   CURVE IS  ATTR CHCE  SPRD IS    ATTR CHCE  PTTRN IS   ATTR CHCE
------------------------    -------------------   -------------------

NF322    ----       ----      ----        ----      ----        ----
MF322    ----       ----      ----        ----      ----        ----

----------------------------------------------------------------------


WRITE OUT DETAILED TEST DATA? (Y OR N)
Y


---------------- SUMMARY OF TOT322 STATISTICAL DATA -----------------

                              GRADE/TIME EQUATION        VARIANCE FROM:
GP.  FILE     GP.       -------------------------     ----------------
NO.  CODE     POP.        (A)        (B)       (R)      GP. EQ.  OWN EQ.
-----------------       -------------------------     ----------------

  TOT322       440      0.80391  0.04137  0.42391      (NA)     0.24109

 1:  NF322     209      0.80396  0.03983  0.41097    0.23304    0.23290
 2:  MF322     231      0.80404  0.04408  0.43393    0.24941    0.24894

----------------------------------------------------------------------


NEXT: FILE WRITE-OUT:

ENTER NO. OF FILE(1-8) OR
9 (NO FILE WANTED)
9
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     RUN COMPLETED. NEXT:
(1) AGAIN (AUTO)? (2) AGAIN (STEP BY STEP)? OR (3) QUIT. (1,2 OR 3)
1
```

```
                                                    Page 3
XXXXXXXXYXXXXXXXXXXXXXXXXXXXXYYXXXXXXXXYXXXXXXXYXXXXXXXXXXXXXXXXXXYXXXXXXX

KEY:
   "    " = NO TEST (N<3)
   "----" = NOT SIG.

---------------------- SUMMARY OF TEST RESULTS------------------------

      ADVANCEMENT CURVES  GRD (SKILL) DISTRIB  OVERALL G/T PATTERN
      ------------------  -------------------  -------------------
FILE  SUBGROUP  P. DIFF.S SUBGROUP  P. DIFF.S  SUBGROUP  P. DIFF.S
CODE  CURVE IS  ATTR CHCE SPRD IS   ATTR CHCE  PTTRN IS  ATTR CHCE
----------------------------------------------------------------- 

NF312   ----     ----      ----      -----      ----      ----
MF312   ----     ----      ----      ----       ----      ----

---------------------------------------------------------------------


--------------- SUMMARY OF TOT312 STATISTICAL DATA -----------------

                            GRADE/TIME EQUATION         VARIANCE FROM:
GP. FILE     GP.        ----------------------------   -----------------
NO. CODE     POP.          (A)       (B)       (R)      GP. EQ.  OWN EQ.
-----------------       ----------------------------   -----------------

  TOT312        68      0.84457   0.06127   0.54180     (NA)     0.32299

 1:  NF312      52      0.84357   0.05050   0.46745    0.31748   0.31391
 2:  MF312      16      0.84175   0.09112   0.66193    0.36326   0.31311

---------------------------------------------------------------------

NEXT: FILE WRITE-OUT:

ENTER NO. OF FILE(1-8) OR
9 (NO FILE WANTED)
9

YXXXXXXXXXXXXXXXXXXXYXXXXXXYXXYXXXXXXXXXXXXXXXXXYXXXXXXXXXXXXXXXXXXXXXX
     RUN COMPLETED. NEXT:
(1) AGAIN (AUTO)? (2) AGAIN (STEP BY STEP)? OR (3) QUIT. (1,2 OR 3)
1

HOW MANY FILES DO YOU WISH TO ANALYZE? (ANS 1-8)
2
PLEASE ENTER CODE NAME OF EACH FILE:
 1=
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX


KEY:
   "    " = NO TEST (N<3)
   "----" = NOT SIG.


---------------------- SUMMARY OF TEST RESULTS ----------------------

       ADVANCEMENT CURVES   GRD (SKILL) DISTRIB   OVERALL G/T PATTERN
      -------------------  --------------------  --------------------
FILE   SUBGROUP  P. DIFF.S  SUBGROUP   P. DIFF.S  SUBGROUP   P. DIFF.S
CODE   CURVE IS  ATTR CHCE  SPRD IS    ATTR CHCE  PTTRN IS   ATTR CHCE
----------------------------  --------------------  --------------------

NF203   ----       ----      ----        ----       ----        ----
MF203   ----       ----      ----        ----       ----        ----

---------------------------------------------------------------------



---------------- SUMMARY OF TOT203 STATISTICAL DATA -----------------

                                    GRADE/TIME EQUATION          VARIANCE FROM:
GP.  FILE        GP.         -------------------------      -----------------
NO.  CODE        POP.          (A)       (B)       (R)        GP. EQ.  OWN EQ.
-----------------            -------------------------      -----------------

  TOT203          463      0.83079  0.09048  0.64413         (NA)     0.38950

 1:  NF203        292      0.83185  0.08644  0.61305       0.40946  0.41047
 2:  MF203        171      0.82916  0.09759  0.69686       0.35762  0.35151

---------------------------------------------------------------------


NEXT: FILE WRITE-OUT:

ENTER NO. OF FILE(1-8) OR
9 (NO FILE WANTED)
9


XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     RUN COMPLETED. NEXT:
(1) AGAIN (AUTO)? (2) AGAIN (STEP BY STEP)? OR (3) QUIT. (1,2 OR 3)
1

HOW MANY FILES DO YOU WISH TO ANALYZE? (ANS 1-8)
2
PLEASE ENTER CODE NAME OF EACH FILE:
 1=
                                                            Page 5
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX


KEY:
   "    " = NO TEST (N<3)
   "----" = NOT SIG.


---------------------- SUMMARY OF TEST RESULTS -----------------------

       ADVANCEMENT CURVES   GRD (SKILL) DISTRIB   OVERALL G/T PATTERN
       --------------------  --------------------  --------------------
FILE   SUBGROUP  P. DIFF.S  SUBGROUP   P. DIFF.S  SUBGROUP   P. DIFF.S
CODE   CURVE IS  ATTR CHCE  SPRD IS    ATTR CHCE  PTTRN IS   ATTR CHCE
---------------------------  --------------------  --------------------

NF318    ----       ----       ----       ----      ----        ----
MF318    ----       ----     SMALLER     <=.05     ----        ----

----------------------------------------------------------------------


---------------- SUMMARY OF TOT318 STATISTICAL DATA ------------------

                          GRADE/TIME EQUATION          VARIANCE FROM:
GP.  FILE     GP.      -------------------------     -----------------
NO.  CODE     POP.       (A)      (B)      (R)        GP. EQ.  OWN EQ.
------------------     -------------------------     -----------------

  TOT318        127    0.78813  0.05653  0.36052       (NA)     0.64457

 1:  NF318       90    0.79043  0.06323  0.38304     0.74782  0.72768
 2:  MF318       37    0.78460  0.03547  0.29402     0.40722  0.36359

----------------------------------------------------------------------


NEXT: FILE WRITE-OUT:

ENTER NO. OF FILE(1-8) OR
9 (NO FILE WANTED)
1
```

```
NF318  FILE:

          0.50  1.50  2.50  3.50  4.50  5.50  6.50  7.50  8.50  9.50
 GRD(WT)
  9.75       0     0     0     0     0     0     0     0     0     0
  9.25       0     0     0     0     0     0     1     0     1     0
  8.75       0     0     0     0     0     0     0     0     0     1
  8.25       0     0     2     0     2     1     2     0     0     0
  7.50       1     0     4     2     1     4     3     1     1     3
  6.50       3     5     7     4     6     5     3     0     0     3
  5.75       2     6     1     2     1     2     3     0     2     0
  5.25       3     1     0     0     0     1     0     0     0     0

NEXT: FILE WRITE-OUT:

ENTER NO. OF FILE(1-8) OR
9 (NO FILE WANTED)
2


MF318  FILE:

          0.50  1.50  2.50  3.50  4.50  5.50  6.50  7.50  8.50  9.50
 GRD(WT)
  9.75       0     0     0     0     0     0     0     0     0     0
  9.25       0     0     0     0     0     0     0     0     0     0
  8.75       0     0     0     0     0     0     0     0     0     0
  8.25       0     0     0     0     0     0     1     0     0     0
  7.50       0     0     0     1     1     1     1     0     0     0
  6.50       2     1     3     1     5     3     3     0     0     0
  5.75       2     2     5     0     0     3     1     1     0     0
  5.25       0     0     0     0     0     0     0     0     0     0

NEXT: FILE WRITE-OUT:

ENTER NO. OF FILE(1-8) OR
9 (NO FILE WANTED)
9

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

```
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX


KEY:
   "    " = NO TEST (N<3)
   "----" = NOT SIG.


---------------------- SUMMARY OF TEST RESULTS -----------------------

       ADVANCEMENT CURVES   GRD (SKILL) DISTRIB   OVERALL G/T PATTERN
       ------------------   -------------------   -------------------
FILE   SUBGROUP  P. DIFF.S  SUBGROUP   P. DIFF.S  SUBGROUP   P. DIFF.S
CODE   CURVE IS  ATTR CHCE  SPRD IS    ATTR CHCE  PTTRN IS   ATTR CHCE
--------------------------  -------------------   -------------------

NF322   LOWER      <=.001    SMALLER    <.0001      ----        ----
NF312   ----        ----     SMALLER    <=.01       ----        ----
NF203   HIGHER     <=.01     SMALLER    <=.01       ----        ----
NF318   ----        ----     GREATER    <=.05       DIFF.      <.0001

---------------------------------------------------------------------


---------------- SUMMARY OF NFCL   STATISTICAL DATA -----------------

                                GRADE/TIME EQUATION        VARIANCE FROM:
GP.  FILE      GP.         -------------------------   -----------------
NO.  CODE      POP.           (A)       (B)      (R)       GP. EQ.  OWN EQ.
------------------         -------------------------   -----------------

  NFCL          643       0.81744  0.07552  0.55855      (NA)     0.54952

 1:  NF322      209       0.80396  0.03683  0.41097     0.37661   0.23290
 2:  NF312       52       0.84357  0.05050  0.46745     0.48096   0.31391
 3:  NF203      292       0.83185  0.08644  0.61305     0.56216   0.41047
 4:  NF318       90       0.79043  0.06323  0.38304     0.97009   0.72768

---------------------------------------------------------------------

NEXT: FILE WRITE-OUT:

ENTER NO. OF FILE(1-8) OR
9 (NO FILE WANTED)
9

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
    RUN COMPLETED. NEXT:
(1) AGAIN (AUTO)? (2) AGAIN (STEP BY STEP)? OR (3) QUIT. (1,2 OR 3)
1
```

```
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

KEY:
   "    " = NO TEST (N<3)
   "----" = NOT SIG.


---------------------- SUMMARY OF TEST RESULTS ----------------------

       ADVANCEMENT CURVES   GRD (SKILL) DISTRIB   OVERALL G/T PATTERN
       ------------------   -------------------   -------------------
FILE   SUBGROUP  P. DIFF.S  SUBGROUP   P. DIFF.S  SUBGROUP   P. DIFF.S
CODE   CURVE IS  ATTR CHCE  SPRD IS    ATTR CHCE  PTTRN IS   ATTR CHCE
----------------------------------------------------------------------

MF322   LOWER     <=.05      SMALLER    <.0001      ----       ----
MF312   ----       ----       ----       ----       ----       ----
MF203   HIGHER    <=.01      SMALLER    <=.01       ----       ----
MF318   LOWER     <=.01       ----       ----       DIFF.     <.0001

----------------------------------------------------------------------


---------------- SUMMARY OF MFCL   STATISTICAL DATA -----------------

                         GRADE/TIME EQUATION         VARIANCE FROM:
GP.  FILE     GP.      -----------------------     ----------------
NO.  CODE     POP.       (A)      (B)      (R)      GP. EQ.  OWN EQ.
-----------------      -----------------------     ----------------

 MFCL          455     0.81271  0.08215  0.62041     (NA)    0.48174

 1:  MF322     231     0.80404  0.04408  0.43393   0.32841   0.24894
 2:  MF312      16     0.84175  0.09112  0.66193   0.64348   0.31311
 3:  MF203     171     0.82916  0.09759  0.69686   0.56163   0.35151
 4:  MF318      37     0.78460  0.03547  0.29402   1.05683   0.36359

----------------------------------------------------------------------

NEXT: FILE WRITE-OUT:

ENTER NO. OF FILE(1-8) OR
9 (NO FILE WANTED)
9


XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

     RUN COMPLETED. NEXT:
(1) AGAIN (AUTO)? (2) AGAIN (STEP BY STEP)? OR (3) QUIT. (1,2 OR 3)
3
```

```
CUMULATED DISTRIBUTION SUM     :
   SUM OF:  TOT322  TOT312  TOT203  TOT318

          0.50  1.50  2.50  3.50  4.50  5.50  6.50  7.50  8.50  9.50
 GRD(WT)
  9.75       0     0     0     0     0     0     0     0     0     0
  9.25       0     0     0     1     2     7     7     1     3     3
  8.75       0     0     0     1     4     4     8     2     4     7
  8.25       1     3     9    13    24    53    21     6     5     3
  7.50      31    42    51    40    48    48    28     4     6     6
  6.50     119   128   109    19    27    23    17     2     2     4
  5.75      82    30    10     2     1     5     4     1     3     0
  5.25      10     3     0     0     0     1     0     0     0     0

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

KEY:
   "    " = NO TEST (N<3)
   "----" = NOT SIG.


---------------------- SUMMARY OF TEST RESULTS ----------------------

       ADVANCEMENT CURVES     GRD (SKILL) DISTRIB    OVERALL G/T PATTERN
       -------------------    -------------------    -------------------
FILE   SUBGROUP  P. DIFF.S    SUBGROUP   P. DIFF.S   SUBGROUP   P. DIFF.S
CODE   CURVE IS  ATTR CHCE    SPRD IS    ATTR CHCE   PTTRN IS   ATTR CHCE
------ -------------------    -------------------    -------------------

TOT322  LOWER       <.0001     SMALLER     <.0001       ----         ----
TOT312   DIFF.      <=.05      SMALLER     <=.01        ----         ----
TOT203  HIGHER      <.0001     SMALLER     <=.001       ----         ----
TOT318  LOWER       <=.01      GREATER     <=.05       DIFF.        <.0001

---------------------------------------------------------------------


WRITE OUT DETAILED TEST DATA? (Y OR N)
Y
---------------- SUMMARY OF SUM    STATISTICAL DATA -----------------

                                GRADE/TIME EQUATION          VARIANCE FROM:
GP.  FILE       GP.      -------------------------      -----------------
NO.  CODE       POP.        (A)       (B)       (R)       GP. EQ.  OWN EQ.
-----------------       -------------------------      -----------------

 SUM            1098     0.81516  0.07871  0.58872      (NA)     0.52173

 1:  TOT322      440     0.80391  0.04137  0.42391    0.34990   0.24109
 2:  TOT312       68     0.84457  0.06127  0.54180    0.52664   0.32299
 3:  TOT203      463     0.83079  0.09048  0.64413    0.56108   0.38950
 4:  TOT318      127     0.78813  0.05653  0.36052    0.98595   0.64457

---------------------------------------------------------------------
```

```
CUMULATED DISTRIBUTION SUM    :
   SUM OF:  TOT322  TOT318

             0.50  1.50  2.50  3.50  4.50  5.50  6.50  7.50  8.50  9.50
 GRD(WT)
  9.75          0     0     0     0     0     0     0     0     0     0
  9.25          0     0     0     0     0     0     1     0     1     0
  8.75          0     0     0     0     0     0     0     0     0     1
  8.25          0     0     2     0     2     1     3     0     0     0
  7.50         15    12    12    13     6    17    10     1     1     3
  6.50         78    94    76    13    21    19    16     1     2     4
  5.75         78    28    10     2     1     5     4     1     3     0
  5.25          7     2     0     0     0     1     0     0     0     0

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX


KEY:
   "    " = NO TEST (N<3)
   "----" = NOT SIG.


---------------------- SUMMARY OF TEST RESULTS ----------------------

        ADVANCEMENT CURVES   GRD (SKILL) DISTRIB   OVERALL G/T PATTERN
        -------------------  --------------------  --------------------
FILE    SUBGROUP  P. DIFF.S  SUBGROUP   P. DIFF.S  SUBGROUP   P. DIFF.S
CODE    CURVE IS  ATTR CHCE  SPRD IS    ATTR CHCE  PTTRN IS   ATTR CHCE
------  -------------------  --------------------  --------------------

TOT322    ----       ----     SMALLER    <=.001       ----        ----
TOT318    ----       ----     GREATER    <.0001      DIFF.      <.0001

----------------------------------------------------------------------



---------------- SUMMARY OF SUM    STATISTICAL DATA -----------------

                                GRADE/TIME EQUATION          VARIANCE FROM:
GP.  FILE        GP.      -------------------------    ----------------
NO.  CODE        POP.       (A)        (B)      (R)      GP. EQ.  OWN EQ.
-----------------------   -------------------------    ----------------

  SUM            567      0.80225  0.04072  0.39879     (NA)     0.33365

 1:  TOT322      440      0.80391  0.04137  0.42391    0.24270   0.24109
 2:  TOT318      127      0.78813  0.05653  0.36052    0.65320   0.64457

----------------------------------------------------------------------
NEXT: FILE WRITE-OUT:

ENTER NO. OF FILE(1-8) OR
9 (NO FILE WANTED)
1
```

```
TOT322 FILE:

                0.50   1.50   2.50   3.50   4.50   5.50   6.50   7.50   8.50   9.50
 GRD(WT)
  9.75            0      0      0      0      0      0      0      0      0      0
  9.25            0      0      0      0      0      0      0      0      0      0
  8.75            0      0      0      0      0      0      0      0      0      0
  8.25            0      0      0      0      0      0      0      0      0      0
  7.50           14     12      8     10      4     12      6      0      0      0
  6.50           73     88     66      8     10     11     10      1      2      1
  5.75           74     20      4      0      0      0      0      0      1      0
  5.25            4      1      0      0      0      0      0      0      0      0

NEXT: FILE WRITE-OUT:

ENTER NO. OF FILE(1-8) OR
9 (NO FILE WANTED)
2


TOT318 FILE:

                0.50   1.50   2.50   3.50   4.50   5.50   6.50   7.50   8.50   9.50
 GRD(WT)
  9.75            0      0      0      0      0      0      0      0      0      0
  9.25            0      0      0      0      0      0      1      0      1      0
  8.75            0      0      0      0      0      0      0      0      0      1
  8.25            0      0      2      0      2      1      3      0      0      0
  7.50            1      0      4      3      2      5      4      1      1      3
  6.50            5      6     10      5     11      8      6      0      0      3
  5.75            4      8      6      2      1      5      4      1      2      0
  5.25            3      1      0      0      0      1      0      0      0      0

NEXT: FILE WRITE-OUT:

ENTER NO. OF FILE(1-8) OR
9 (NO FILE WANTED)
9

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
     RUN COMPLETED. NEXT:
(1) AGAIN (AUTO)? (2) AGAIN (STEP BY STEP)? OR (3) QUIT. (1,2 OR 3)
1
HOW MANY FILES DO YOU WISH TO ANALYZE? (ANS 1-8)
2
PLEASE ENTER CODE NAME OF EACH FILE:
 1=
NFCL
 2=
MFCL
```

```
                                                        Page 12
CUMULATED DISTRIBUTION SUM    :
   SUM OF:  NFCL     MFCL

          0.50  1.50  2.50  3.50  4.50  5.50  6.50  7.50  8.50  9.50
 GRD(WT)
  9.75       0     0     0     0     0     0     0     0     0     0
  9.25       0     0     0     1     2     7     7     1     3     3
  8.75       0     0     0     1     4     4     8     2     4     7
  8.25       1     3     9    13    24    53    21     6     5     3
  7.50      31    42    51    40    48    48    28     4     6     6
  6.50     119   128   109    19    27    23    17     2     2     4
  5.75      32    30    10     2     1     5     4     1     3     0
  5.25      10     3     0     0     0     1     0     0     0     0


XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX


KEY:
   "    " = NO TEST (N<3)
   "----" = NOT SIG.


---------------------- SUMMARY OF TEST RESULTS ----------------------

       ADVANCEMENT CURVES   GRD (SKILL) DISTRIB   OVERALL G/T PATTERN
      -------------------- --------------------  --------------------
FILE   SUBGROUP  P. DIFF.S  SUBGROUP   P. DIFF.S  SUBGROUP   P. DIFF.S
CODE   CURVE IS  ATTR CHCE  SPRD IS    ATTR CHCE  PTTRN IS   ATTR CHCE
------------------------- --------------------  --------------------

NFCL     ----       ----      ----       ----       ----       ----
MFCL     ----       ----      ----       ----       ----       ----

----------------------------------------------------------------------

--------------- SUMMARY OF SUM    STATISTICAL DATA ------------------

                           GRADE/TIME EQUATION          VARIANCE FROM:
GP.  FILE     GP.     -------------------------      -----------------
NO.  CODE     POP.        (A)      (B)      (R)         GP. EQ.  OWN EQ.
-----------------     -------------------------      -----------------

  SUM         1098    0.81516  0.07871  0.58872       (NA)     0.52173

 1:  NFCL      643    0.81744  0.07552  0.55855      0.55048   0.54952
 2:  MFCL      455    0.81271  0.08215  0.62041      0.48222   0.48174

----------------------------------------------------------------------

NEXT: FILE WRITE-OUT:

ENTER NO. OF FILE(1-8) OR
9 (NO FILE WANTED)
9
```

APPENDIX G

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

HIREST

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

HIREST

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

HIREST TECHNICAL ANALYSIS

"HIREST" combines several different types of techniques to achieve its purpose - the estimating of group losses and hires over a multi-year period.

The first section is concerned with the estimation of loss probabilities for the employees whose data are contained in the file input at the start of the program. The methods used for this purpose are as follows:

1. Subroutine "LOSS" computes the probability of loss due to causes other than death, disability or retirement by means of the group L-P equation.

The "L-P"equation referred to is the log-probability turnover curve equation of the group being analyzed. For a detailed explanation of such equations,see the Technical Analysis for program LOGPRO. As indicated there, L-P equations are of the form

$$q(Y) = A + B(\text{LOG } X)$$

LOSS determines from the input A and B values the probability of retention from point T_1 to T_2 by computing the value of the equation for points $X = T_1$ and $X = T_2$, and dividing the value at T_2 by the value at T_1 . Subtracting this quotient from 1 gives the net L-P loss probability for the individual for the T_1 - to - T_2 time period.

2. Subroutine DEATH determines, by reference to the employee's sex and age, the probability of death during the year involved.

Values for this subprogram are taken from mortality tables of the USCSC Chief Actuary. A semi-log plot of table values indicates that when the tabled probabilities are converted to logarithms their values are approximately linear functions of employee age. Accordingly a linear curve of form

$$\log Y = A + BX$$

was fitted to the tabled data and the resulting equation is used in the subroutine in place of a more cumbersome look-up table program.

3. Subroutine DISAB computes the probability of loss due to disability by means similar to those used in DEATH, above.

In these two applications, the equation fits are reliable to two decimal places, which is sufficient precision for use in small-to-moderate-size employee groups expected for HIREST.

4. Subroutine RETIRE determines the probability of retirement loss.

In this subroutine, the normally more cumbersome look-up table technique is used because of the need for testing for several categories of retirement eligibility.

The loss rates shown are directly applicable only to regular employees of the U.S. Civil Service. Use of HIREST on other populations will require development of a new or revised RETIRE subprogram giving rates appropriate for such other jurisdictions or retirement systems.

5. The sum of the loss probabilities is subtracted from 1 to obtain the employee's retention probability for the time period specified.

There is one exception to step 5: in the base period, the period from the as-of date of the employee file to the start of the first fiscal year of the projection period, the annual rates derived in steps 2-4 are first multiplied by BASE, the length of the base period in years.

In this we are adopting the standard actuarial assumption that such annual probabilities are approximately linear functions of the time of exposure. Thus, for example, if the one-year probability of a particular kind of loss for a person of given age and sex is .012, we assume that the probability of such loss over, say, only a five-month base period is 5/12ths of the annual rate or .005.

Such linear interpolation is not needed for step 1 because the L-P equation technique computes the probability for the exact base period. Such interpolation is also not needed in the projection period because the time intervals involved are all full years.

The second major section of HIREST is concerned with estimating the numbers and types of employees to be added to the starting workforce during each year of the projection period and estimating losses among these new hires during the projection period.

This process is essentially one of iterations. Given the number of employees in file at the start of the base period (i.e., the as-of date of the employee data file), HIREST first estimates the number of such employees who can expect to be lost during the base period and then subtracts the remaining employees from the total employment figure specified for the start of the projection period to obtain the number of additional employees who will have to be on board.

The number of additional employees needed on board by the end of the base period, however, is not quite the total number of new employees that need to be hired during the base period. This is because some of the people that will need to be hired during the base period can be expected to leave before end of the base period.

For example, suppose the base period is one year long. Assuming that new hiring is evenly distributed over that year, the new hires still on board at the end of the base period will have an average length of service of 0.5 year. Suppose further that the group L-P equation shows that only 90% of new hires into the group are still left at the end of 0.5 years from date of hire. Obviously, then, if we need to have 9 new employees on board at the end of the base period, our total base period hiring will have to be 10, not 9.

Thus HIREST first estimates the number of additional employees needed on board at the end of the period and then, using a retention rate equivalent to half the base period, estimates the number of new hires needed. For that fraction of total accessions who are expected to be new hires, the retention rate is computed from the group L-P equation. For that fraction who are expected to be accessions to the group from elsewhere in the Service, the retention rate is one-half the average group rate for the period involved.

With the number of additional end-of-year employees left from (a) new hires and (b) accessions estimated, HIREST then estimates losses in these two groups during each subsequent period of the projection. The "new hires" employee retention over subsequent years is estimated from the L-P curve (for 0.5 to 1.5 years, 1.5 to 2.5 years, etc.). Retention for "accessions" employees is assumed to be the same as for the retained original group population as a whole.

This process is then repeated to estimate the number of new employees needed for each subsequent year of the projections.

Several additional points should be made concerning specific features of HIREST and the methodological assumptions that it reflects.

First, HIREST asks for input of the proportion of total accessions who will be outside hires and uses this figure for all projection years without change because it is our experience that the proportion of outside hires to other accessions in most occupations is the result of continuing staffing policy and practice, and as such, tends to hold constant over time regardless of changes in total employment trend.

Second, as indicated earlier, HIREST assumes that loss rates among new accessions from elsewhere in the service will be on a par, for practical purposes, with the loss rate among continuing employees. In some cases, there may be some difference in total rate to be noted, depending on the sources from which in-service accessions are drawn, or in the relative distribution of losses by type (proportion of retirements, etc.). In most cases, we would expect these differences to be minor except (a) where the existing workforce contains a strongly disproportionate share of retirement eligibles or (b) where the proportion of total accessions who are new hires is unusually low. In such cases, the absolute level of losses estimated can be expected to vary somewhat from actual experience. The relative effects of changing projection parameters, however, can be expected to be reliable (turnover goes up when employment rises, etc.)

Third, the RIF subsection employs a simplified method of RIF-effect estimation which does not differentially reduce the population in "Last-Hired, First-Out" order, since RIF situations frequently vary in this respect. As a result, the loss rates estimated by HIREST may not show as much of a drop in the first period(s) following a major RIF as may actually be likely to occur. In most cases, this difference is unlikely to be very large. Correction for this special case, however, would require additional program features and complexities which are beyond the scope of a basic-type program.

And finally, HIREST projections show only the projected "expected values" - i.e., HIREST does not compute the probable range of reliability of expected values, as does the LOGPRO program. This is because HIREST is a hybrid program using both probabilistic methods (L-P technique) and deterministic techniques (the actuarial-table subroutines). Since the probable error range of the actuarial components is not known, the probable error of the resulting overall projections cannot be readily determined.

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

HIREST

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

*= COMPUTER WRITTEN

"HIREST" IS A FORTRAN IV PROGRAM FOR ESTIMATING LOSS RATES AND HIRING NEEDS IN A SPECIFIC WORK GROUP OVER A 1-5 YEAR PERIOD. IT REQUIRES AN INPUT FILE OF DATA ON EACH GROUP EMPLOYEE (DOB, EOD,SEX,SCD) AND ENTRY OF (1) THE NUMBER OF EMPLOYEES IN THE FILE,(2) THE GROUP LOG-PROBABILITY LOSS EQUATION, (3) THE EXPECTED GROUP POPULATION DURING EACH YEAR OF THE PROJECTION PERIOD, AND (4) THE PROPORTION OF TOTAL GROUP ACCESSIONS WHO ARE EXPECTED TO BE NEW HIRES.

REQUIRED SUBPROGRAMS ARE DATE, LOSS, DEATH, DISAB, RETIRE AND ANDXP.

EXECUTION COMMANDS

COMMANDS FOR (A) TRANSLATING THE MAIN PROGRAM AND REQUIRED SUBPROGRAMS INTO MACHINE LANGUAGE (I.E., THE "COMPILATION" PHASE) AND (B) LOADING THE COMPILED PROGRAM AND SUBPROGRAMS INTO THE CENTRAL PROCESSING AREA AND STARTING THE RUN (I.E., THE "LOADING AND EXECUTION" PHASE) VARY ACCORDING TO THE COMPUTER SYSTEM BEING USED.

THE "HIREST" RUN BEGINS:

```
* ENTER 5-SPACE NAME OF
* EMPLOYEE DATA FILE
```

THE OPERATOR RESPONDS BY ENTERING THE CODE NAME OF THE GROUP EMPLOYEE DATA FILE (E.G., SAMPLE DATA FILE MP801 OR MP322). (NOTE THAT ALL DATA ENTRIES ARE FOLLOWED BY A CARRIAGE-RETURN.)

```
  MP801
   OR
  MP322
```

NEXT THE COMPUTER ASKS:

```
* ENTER NO. OF EMPLOYEES
* IN (MP801) FILE
```

ENTER THE NUMBER IN INTEGER FORM (NO DECIMAL POINT) AND FOLLOW WITH A CR.

NEXT THE COMPUTER ASKS:

```
* ENTER (MP801) AS-OF (I.E.,
* CURRENT) DATE IN MO., YR. (E.G.,05,75)
```

ENTER THE DATE OF THE FILE CALLED PREVIOUSLY. USE INTEGER FORMAT: E.G., MAY 1975 = "05,75".

(NOTE THAT ALL DATES IN THIS PROGRAM AND ITS FILES ARE EXPRESSED IN THIS WAY. INDIVIDUAL DAYS ARE IGNORED.)

* = COMPUTER WRITTEN

NEXT:

```
* ENTER A, B OF (MP801) L-P EQUATION
```

THIS REFERS TO THE GROUP LOG-PROBABILITY LOSS EQUATION (SEE DOCUMENTATION FOR COMPUTER PROGRAM "LOGPRO"). SUCH EQUATIONS ARE OF THE FORM F(Y)=A+B(LOG X). IT IS THE A AND B COEFFICIENTS WHICH ARE NEEDED HERE. DO NOT FORGET TO ENTER THE NEGATIVE SIGN FOR THE B-VALUE. SHOW DECIMAL POINTS. E.G.:

```
  .9022,-1.
```

NEXT THE COMPUTER ASKS:

```
* ENTER NO. OF FISCAL YEARS
* PROJECTION WANTED (1-5)
```

ENTER THE NUMBER DESIRED IN INTEGER FORM; E.G.,

```
  5
```

NEXT THE COMPUTER ASKS FOR INPUT OF THE DATA ITEMS NEEDED TO MAKE THE DESIRED PROJECTION.

```
* ENTER MO., YEAR OF START
* OF FIRST FISCAL YEAR (E.G., 10,76)
```

SELF EXPLANATORY. OCTOBER 1976 = "10,76".

```
* ENTER (MP801) POPULATION
* AT START OF FY (1976):
```

ENTER POPULATION DATA IN INTEGER FORMAT. THE COMPUTER CONTINUES:

```
* ENTER (MP801) POPULATION
* AT END OF FISCAL YEAR
*
* 1976
```

ENTER DATA INDICATED. COMPUTER CONTINUES TO REQUEST END-OF-YEAR POPULATION DATA FOR EACH OF THE YEARS OF THE PROJECTION.

NEXT THE COMPUTER ASKS:

```
* ENTER FRACTION OF (MP801) TOTAL
* ACCESSIONS WHO ARE NEW HIRES (E.G., 0.25)
```

ENTER THE INDICATED DATA, BEING SURE TO SHOW THE NECESSARY DECIMAL POINT.

FROM THIS POINT ON THE PROGRAM RUNS AUTOMATICALLY , PERFORMING THE NECESSARY COMPUTATIONS AND PRINTING OUT THE RESULTS IN TABULAR FORM. THE COMPUTER THEN ASKS:

```
* = COMPUTER WRITTEN

* RUN AGAIN? (Y OR N)
```

IF NO FURTHER RUN IS NEEDED, ENTER "N". IF A "Y" IS ENTERED, THE COMPUTER THEN ASKS:

```
* (1) FROM TOP; OR (2) WITH NEW EMP.T DATA
* (1 OR 2)
```

IF "1" IS ENTERED THE PROGRAM RECYCLES TO THE FIRST STEP. IF "2" IS ENTERED THE RECYCLE IS TO THE QUESTION:

```
* ENTER NO. OF FISCAL YEARS
* PROJECTION WANTED (1-5)
```

STAFFING NEEDS PLANNING COMPUTER PROGRAM:

HIREST

BUREAU OF POLICIES AND STANDARDS
UNITED STATES CIVIL SERVICE COMMISSION
WASHINGTON, D.C. 20415

```
C       MAIN PROGRAM HIREST. INPUTS FILE OF EMPLOYEE DATA
C       (DOB,EOD,SEX,SCD).  REQUIRES ENTRY OF GROUP L-P
C       EQUATION, FUTURE EMPLOYMENT LEVELS (1-5 YEARS), AND
C       PROPORTION OF TOTAL ACCESSIONS WHO ARE NEW HIRES.
C       PROJECTS LOSSES DUE TO L-P CURVE, DEATH, DISABILITY,
C       AND RETIREMENTS; ESTIMATES RIF LOSSES; ESTIMATES
C       NEW HIRES AND ACCESSIONS.  REQUIRED SUBROUTINES:
C       DATE, LOSS, DEATH, DISAB, RETIRE, ANDXP. (3/75)
        DIMENSION POP(7,14),MPREC(200,13),SUML(6),SUMD(6),
        1  SUMDI(6),SUMR(6),SUMSEP(6),HL(6),HIR(6),
        2  ACC(6),SUMH(6),PCT(6,6),R(6),RIF(6),IYR(5)
40      WRITE(5,50)
50      FORMAT(//,1X,'ENTER 5-SPACE NAME OF ',/,
        1  1X,'EMPLOYEE DATA FILE',/)
        READ(5,80)NFILE
80      FORMAT(A5)
        CALL IFILE(20,NFILE)
        WRITE(5,110)NFILE
110     FORMAT(//,1X,'ENTER NO. OF EMPLOYEES',/,
        1  1X,'IN ',A5,' FILE',//)
        READ(5,140)NEMP
140     FORMAT(I4)
        WRITE(5,170)NFILE
170     FORMAT(//,1X,'ENTER ',A5,' AS-OF (I.E.,',/,
        1  1X,'CURRENT) DATE IN MO., YR. (E.G.,05,75)',/)
        READ(5,*)MM,MYR
        WRITE(5,210)NFILE
210     FORMAT(//,1X,'ENTER A,B OF ',A5,' L-P EQUATION',/)
        READ(5,*)A,B
        DO 245 I=1,NEMP
        READ(20,240)(MPREC(I,J),J=1,13)
240     FORMAT(13I1)
245     CONTINUE
        WRITE(5,260)NFILE
260     FORMAT(//,1X,'FILE ',A5,' READ. NEXT:',//)
270     WRITE(5,280)
280     FORMAT(1X,'ENTER NO. OF FISCAL YEARS',/,
        1  1X,'PROJECTION WANTED (1-5):',/)
        READ(5,310)NYPRO
310     FORMAT(I1)
        NT=NYPRO+1
        NPOP=NYPRO+2
        WRITE(5,340)
340     FORMAT(/,1X,'ENTER MO., YEAR OF START')
        WRITE(5,350)
350     FORMAT(1X,'OF FIRST FISCAL YEAR (E.G., 10,76)',/)
360     READ(5,*)MO,NYR
        IYR(1)=1901+NYR
        IF(NYPRO.EQ.1) GO TO 430
        DO 420 I=2,NYPRO
        J=I-1
        IYR(I)=IYR(J)+1
420     CONTINUE
430     DO 460 I=1,7
```

```
      DO 460 J=1,14
      POP(I,J)=0.0
460   CONTINUE
      POP(1,14)=NEMP
      POP(1,13)=NEMP
      CALL DATE(MO,NYR,MM,MYR,BASE)
480   WRITE(5,490)NFILE,IYR(1)
490   FORMAT(//,1X,'ENTER ',A5,' POPULATION',/,
      1  1X,'AT START OF FY ',I4,':',/)
      READ(5,520)NST
520   FORMAT(I4)
      POP(2,14)=NST
      WRITE(5,550)NFILE
550   FORMAT(//,1X,'ENTER ',A5,' POPULATION',/,
      1  1X,'AT END OF FISCAL YEAR')
      DO 630 I=1,NYPRO
      J=I+2
      WRITE(5,600)IYR(I)
600   FORMAT(/,1X,I4,':',/)
      READ(5,620)NEND
620   FORMAT(I4)
630   POP(J,14)=NEND
640   WRITE(5,650)NFILE
650   FORMAT(//,1X,'ENTER FRACTION OF ',A5,' TOTAL',/,
      1  1X,'ACCESSIONS WHO ARE NEW HIRES (E.G.,0.25):',/)
      READ(5,680)PCTNU
680   FORMAT(F5.3)
      DO 760 I=1,6
      SUML(I)=0.0
      RIF(I)=0.0
      SUMD(I)=0.0
      SUMDI(I)=0.0
      SUMR(I)=0.0
      SUMSEP(I)=0.0
760   CONTINUE
      DO 780 I=1,6
780   HL(I)=0.0
      DO 1170 I=1,NEMP
      MDOB=10*MPREC(I,1)+MPREC(I,2)
      NDOB=10*MPREC(I,3)+MPREC(I,4)
      CALL DATE(MO,NYR,MDOB,NDOB,AGE)
      MTIM=10*MPREC(I,5)+MPREC(I,6)
      NTIM=10*MPREC(I,7)+MPREC(I,8)
      CALL DATE(MM,MYR,MTIM,NTIM,T1)
      CALL DATE(MO,NYR,MTIM,NTIM,T2)
      MSCD=10*MPREC(I,10)+MPREC(I,11)
      NSCD=10*MPREC(I,12)+MPREC(I,13)
      CALL DATE(MO,NYR,MSCD,NSCD,SERV)
      SEX=MPREC(I,9)
      DO 1170 J=1,NT
      CALL LOSS(A,B,T1,T2,PL)
      CALL DEATH(AGE,SEX,PD)
      CALL DISAB(AGE,SEX,PDI)
      CALL RETIRE(AGE,SEX,SERV,PR)
      IF((BASE.EQ.0.0).OR.(J.NE.1)) GO TO 1000
      PD=PD*BASE
      PDI=PDI*BASE
      PR=PR*BASE
1000  TOT=PL+PD+PDI+PR
      IF(TOT.GT.1.) TOT=1.
      R(J)=1.-TOT
      IF(J.EQ.1) GO TO 1060
      K=J-1
      R(J)=R(J)*R(K)
      RK=R(K)
1060  L=J+1
      IF(J.EQ.1)RK=1.
      POP(L,13)=POP(L,13)+R(J)
      SUML(J)=SUML(J)+PL*RK
      SUMD(J)=SUMD(J)+PD*RK
      SUMDI(J)=SUMDI(J)+PDI*RK
      IF(TOT.EQ.1.)PR=1.-(PL+PD+PDI)
      SUMR(J)=SUMR(J)+PR*RK
      T1=T2
      T2=T1+1.
      AGE=AGE+1.
      SERV=SERV+1.
1170  CONTINUE
      BMULT=1./ANDXP(A+B*ALOG10(BASE/2.))
      BINP=POP(2,14)-POP(2,13)
      IF(BINP)1260,1210,1210
1210  HIR(1)=BINP*PCTNU*BMULT
      ACC(1)=BINP*((1.+POP(2,13)/POP(1,13))/2.)*(1.-PCTNU)
      POP(2,1)=BINP*PCTNU
      POP(2,2)=BINP-POP(2,1)
      HL(1)=(HIR(1)-POP(2,1))+(ACC(1)-POP(2,2))
      GO TO 1270
1260  RIF(1)=ABS(BINP)
      DO 1263 LO=2,NPOP
      POP(LO,13)=POP(LO,13)-RIF(1)
1263  CONTINUE
      SUML(1)=SUML(1)+RIF(1)
1270  T1=BASE
      J=3
      DO 1370 I=3,NPOP
      T2=T1+1.
      CALL LOSS(A,B,T1,T2,PL)
      K=I-1
      L=J-2
      PL=1.-PL
      POP(I,J)=POP(K,L)*PL
      T1=T2
      J=J+2
1370  CONTINUE
      J=4
      DO 1450 I=3,NPOP
      K=I-1
      L=J-2
      RATE=POP(I,13)/POP(K,13)
      POP(I,J)=POP(K,L)*RATE
```

```
          J=J+2
1450      CONTINUE
          BMULT=1./ANDXP(A+B*ALOG10(0.5))
          DO 1796 IY=3,NPOP
          T1=0.5
          LH=IY-1
          SUM=0.0
          DO 1530 LSUM=3,13
          SUM=SUM+POP(IY,LSUM)
1530      CONTINUE
          BINP=POP(IY,14)-SUM
          IF(BINP)1785,1782,1550
1550      HIR(LH)=BINP*PCTNU*BMULT
          ACC(LH)=BINP*((1.+POP(IY,13)/POP(LH,13))/2.)*(1.-PCTNU)
          POP(IY,1)=BINP*PCTNU
          POP(IY,2)=BINP-POP(IY,1)
          HL(LH)=(HIR(LH)-POP(IY,1))+(ACC(LH)-POP(IY,2))
          IF(IY.EQ.NPOP) GO TO 1796
          T1=0.5
          IJ=IY+1
          J=3
          DO 1710 I=IJ,NPOP
          T2=T1+1.
          CALL LOSS(A,B,T1,T2,PL)
          PR=1.-PL
          K=I-1
          L=J-2
          POP(I,J)=POP(K,L)*PR
          T1=T2
          J=J+2
1710      CONTINUE
          J=4
          DO 1780 I=IJ,NPOP
          K=I-1
          L=J-2
          RATE=POP(I,13)/POP(K,13)
          POP(I,J)=POP(K,L)*RATE
          J=J+2
1780      CONTINUE
          GO TO 1796
1782      HIR(LH)=0.0
          ACC(LH)=0.0
          GO TO 1796
1785      RIF(LH)=ABS(BINP)
          HIR(LH)=0.0
          ACC(LH)=0.0
          SUML(LH)=SUML(LH)+RIF(LH)
          DO 1795 LO=IY,NPOP
          BA=POP(IY,13)
          POP(LO,13)=POP(LO,13)-RIF(LH)*(POP(LO,13)/BA)
1795      CONTINUE
1796      CONTINUE
```

```
      DO 2170 I=1,NT
      SUMSEP(I)=SUML(I)+SUMD(I)+SUMDI(I)+SUMR(I)
      IF(I.EQ.1) GO TO 2110
      SUML(I)=SUML(I)/SUMSEP(I)
      SUMD(I)=SUMD(I)/SUMSEP(I)
      SUMDI(I)=SUMDI(I)/SUMSEP(I)
      SUMR(I)=SUMR(I)/SUMSEP(I)
      J=I+1
      SUM=0.0
      DO 1910 L=2,10,2
      M=L+2
      SUM=SUM+(POP(I,L)-POP(J,M))
1910  CONTINUE
      SUM=SUM+(POP(I,13)-POP(J,13))
      SUM=SUM+(ACC(I)-POP(J,2))
      SUML(I)=SUML(I)*SUM
      SUMD(I)=SUMD(I)*SUM
      SUMDI(I)=SUMDI(I)*SUM
      SUMR(I)=SUMR(I)*SUM
      SUMM=0.0
      DO 2020 L=1,9,2
      M=L+2
      SUMM=SUMM+(POP(I,L)-POP(J,M))
2020  CONTINUE
      SUML(I)=SUML(I)+SUMM+(HIR(I)-POP(J,1))
      SUMSEP(I)=SUML(I)+SUMD(I)+SUMDI(I)+SUMR(I)
2110  EN=POP(I,14)
      PCT(I,1)=(SUML(I)/EN)*100.
      PCT(I,2)=(SUMD(I)/EN)*100.
      PCT(I,3)=(SUMDI(I)/EN)*100.
      PCT(I,4)=(SUMR(I)/EN)*100.
      PCT(I,5)=(SUMSEP(I)/EN)*100.
2170  CONTINUE
      DO 2200 I=1,NT
      SUMH(I)=HIR(I)+ACC(I)
2200  CONTINUE
      WRITE(5,2220)
2220  FORMAT(///,17X,'SUMMARY OF ESTIMATED LOSSES AND GAINS')
      WRITE(5,2240)(IYR(I),I=1,NYPRO)
2240  FORMAT(//,1X,'PERIOD:',6X,'(BASE)      ',I4,4(6X,I4))
      WRITE(5,2260)
2260  FORMAT(/,1X,'POPULATION:')
      WRITE(5,2280)(POP(I,14),I=1,NT)
2280  FORMAT(/,3X,'START',1X,6(6X,F4.0))
      WRITE(5,2300)(POP(I,14),I=2,NPOP)
2300  FORMAT(3X,'END    ',6(6X,F4.0))
      WRITE(5,2320)
2320  FORMAT(/,1X,'EST. LOSSES',6(' NO.   %  '))
      WRITE(5,2340)((SUML(I),PCT(I,1)),I=1,NT)
2340  FORMAT(/,3X,'LOSS *  ',6(F5.0,F5.1))
      WRITE(5,2360)((SUMD(I),PCT(I,2)),I=1,NT)
2360  FORMAT(3X,'DEATH   ',6(F5.0,F5.1))
```

```
      WRITE(5,2380)((SUMDI(I),PCT(I,3)),I=1,NT)
2380  FORMAT(3X,'DISAB.    ',6(F5.0,F5.1))
      WRITE(5,2400)((SUMR(I),PCT(I,4)),I=1,NT)
2400  FORMAT(3X,'RETIRE    ',6(F5.0,F5.1))
      WRITE(5,2420)((SUMSEP(I),PCT(I,5)),I=1,NT)
2420  FORMAT(/,1X,'TOTAL        ',6(F5.0,F5.1))
      WRITE(5,2432)(RIF(I),I=1,NT)
2432  FORMAT(/,3X,'*(RIF)    ',6(F5.0,5X))
      WRITE(5,2440)
2440  FORMAT(/,1X,'EST. GAINS',/)
      WRITE(5,2460)(HIR(I),I=1,NT)
2460  FORMAT(3X,'NEW HIRES ',6(F5.0,5X))
      WRITE(5,2480)(ACC(I),I=1,NT)
2480  FORMAT(3X,'ACCESSIONS',6(F5.0,5X))
      WRITE(5,2500)(SUMH(I),I=1,NT)
2500  FORMAT(/,1X,'TOT. GAINS   ',6(F5.0,5X))
      WRITE(5,2520)
2520  FORMAT(///,1X,'RUN AGAIN? (Y OR N)',/)
      READ(5,2540)IX
2540  FORMAT(A1)
      IF(IX.EQ.'N') GO TO 2620
      WRITE(5,2570)
2570  FORMAT(/,1X,'(1) FROM TOP; OR (2) WITH NEW EMP.T DATA',/
      1  ,1X,'(1 OR 2)',/)
      READ (5,2590)IZ
2590  FORMAT(I1)
      IF(IZ.EQ.1) GO TO 40
      GO TO 270
2620  STOP
      END
      SUBROUTINE DATE(MOI,MYI,MOF,MYF,DAT)
      NY=MYI-MYF
      NM=MOI-MOF
      EN=NM
      EN=EN/12.
      ENY=NY
      DAT=ENY+EN
      RETURN
      END
      SUBROUTINE LOSS(A,B,T1,T2,PR)
      PT1=ANDXP(A+B*ALOG10(T1))
      PT2=ANDXP(A+B*ALOG10(T2))
      PR=1.-PT2/PT1
      RETURN
      END
```

```
          SUBROUTINE DEATH(AGE,SEX,PD)
          IF(SEX.EQ.1.)GO TO 50
          IF(AGE.GT.60.)GO TO 40
          PD=10.**(-4.07789+.03503*AGE)
          GO TO 60
 40       PD=10.**(-4.443573+.040977*AGE)
          GO TO 60
 50       IF(AGE.GT.60.)GO TO 55
          PD=10.**(-3.98803+.02685*AGE)
          GO TO 60
 55       PD=10.**(-5.490235+.05113*AGE)
 60       RETURN
          END
```

```
      SUBROUTINE DISAB(AGE,SEX,PDI)
      IF(SEX.EQ.1.) GO TO 50
      PDI=10.**(-10.25073+4.87154*ALOG10(AGE))
      GO TO 60
50    PDI=10.**(-9.05390+4.19356*ALOG10(AGE))
60    RETURN
      END
      SUBROUTINE RETIRE(AGE,SEX,SERV,PRET)
      DIMENSION RML(13),RFL(13),RMM(16),RFM(16)
      DATA (RML(I),I=1,13)/.06,.05,.06,.11,.09,.09,.1,.1,.11,
      1  .12,.13,.14,1./,(RMM(I),I=1,16)/.25,.16,.15,.15,.16,
      2  .21,.19,.18,.18,.19,.23,.22,.22,.23,.62,1./,(RFL(I),
      3  I=1,13)/.08,.06,.07,.12,.1,.1,.13,.14,.15,.16,.17,
      4  .18,1./,(RFM(I),I=1,16)/.29,.2,.18,.18,.19,.26,.2,
      5  .22,.19,.18,.23,.21,.2,.22,.61,1./
      IF(((AGE.GE.55.).AND.(AGE.LT.60.)).AND.(SERV.GE.30.))
      1  GO TO 150
      IF(((AGE.EQ.60.).OR.(AGE.EQ.61.)).AND.(SERV.GE.15.))
      1  GO TO 200
      IF((AGE.GE.62.).AND.(SERV.GE.5.))GO TO 250
      PRET=0.0
      GO TO 340
150   YR=AGE-54.
      IYR=YR
      IF(SEX.EQ.2.)PRET=RMM(IYR)
      IF(SEX.EQ.1.)PRET=RFM(IYR)
      GO TO 340
200   YR=AGE-54.
      IYR=YR
      IF(SEX.EQ.2.)PRET=RMM(IYR)
      IF(SEX.EQ.1.)PRET=RFM(IYR)
      GO TO 340
250   IF(SERV.GE.12.)YR=AGE-54.
      IF(SERV.LT.12.)YR=AGE-61.
      IYR=YR
      IF(SEX.EQ.1.) GO TO 320
      IF(SERV.GE.12.)PRET=RMM(IYR)
      IF(SERV.LT.12.)PRET=RML(IYR)
      GO TO 340
320   IF(SERV.GE.12.)PRET=RFM(IYR)
      IF(SERV.LT.12.)PRET=RFL(IYR)
340   RETURN
      END
```

```
          FUNCTION ANDXP(X)
C         ACCUMULATIVE NORMAL DISTRIBUTION FUNCTION. APPROXIMATES P
C         FROM X (TO + OR - .0000001). (NBS-55, P. 932)
          X1 = ABS(X)
          T = 1./(1.+.2316419*X1)
          ANDXP = 1.-.3989423*EXP(-(X1**2)/2.)*(.3193815*T-.3565638*T**2
          1  +1.781478*T**3-1.821256*T**4+1.330274*T**5)
          IF (X.LT.0.0) ANDXP = 1.-ANDXP
          RETURN
          END
```

APPENDIX H

RUNNING THE COMPUTER PROGRAMS

326

APPENDIX H

RUNNING THE COMPUTER PROGRAMS

The programs described in this handbook are all written in FORTRAN IV and are designed to be run on a time sharing computer system. As much as possible these programs are written in standard straightforward FORTRAN which should adapt to any time-sharing system. Despite this, however, there will probably be system differences which will require a user to make some editorial changes in the programs.

There are two points in the process of entering and running a program into the computer where such differences will become known. The following steps give a general outline of the process and show where system differences may pop up.

1. Enter the program into the computer line by line (from a Program Listing).
2. Store the program under its given name (using a command such as "SAVE").
3. Compile the program. The command for this will vary from system to system and will, in many cases, cause the computer to print out compilation error messages which will be the result of system differences.
4. If compilation error messages occur, use the system's editing procedures to change the program lines causing the messages. Repeat steps 3 and 4 until there are no more compilation messages.
5. Run the program. Here again, commands for this step vary from system to system and what are known as execution error messages may occur. These messages mean that further editing of the program will be needed.
6. Upon the elimination of the execution error messages, the program should run and give the desired analysis results.

Some of the areas more likely to need changing as a result of error messages are:

1. Input/Output unit numbers.
2. The process of calling up data files.
3. The symbols used for continuation and comment lines.

4. Control characters in FORMAT statements.

5. Line format.

6. Use of specification statements.

All main programs and subroutines are entered and stored separately. At run time the method of calling up a set of main program plus subroutines will depend on the rules of the time-sharing system being used.

When running, each main program will ask to be given data. The form of the responses to these questions is discussed both in the chapter pertaining to the program and in each program's Operation Manual.

Although the programs are written for a time-sharing environment, they may be adapted for use in batch systems.

If further assistance is needed either in the running of these programs or in revising the programs to adapt to a particular use, contact the address in the Preface.

U.S. GOVERNMENT PRINTING OFFICE : 1977 O—247-508